Causal Approximations

PANDU NAYAK
RECOM TECHNOLOGIES
ARTIFICIAL INTELLIGENCE RESEARCH BRANCH
MAIL STOP 269-2
NASA AMES RESEARCH CENTER
MOFFETT FIELD, CA 94035-1000

NASA Ames Research Center
Artificial Intelligence Research Branch

Technical Report FIA-93-13

May, 1993

Causal Approximations

P. Pandurang Nayak
Recom Technologies, NASA Ames Research Center
AI Research Branch
Mail Stop 269-2
Moffett Field, CA 94035.
Email: nayak@ptolemy.arc.nasa.gov

Abstract

Adequate problem representations require the identification of abstractions and approximations that are well suited to the task at hand. Constructing such representations is often difficult. In this paper we analyze the problem of automatically selecting adequate models for the task of generating *parsimonious causal explanations*. This paper makes three important contributions. First, it develops a precise formalization of the problem of finding models that generate parsimonious causal explanations. In this formalization, models are defined as sets of *model fragments*, causal explanations are generated using *causal ordering*, and model simplicity is based on the intuition that using more approximate descriptions of fewer phenomena leads to simpler models. Second, it uses this formalization to show that, in general, the problem is intractable. In addition, it identifies three sources of intractability: (a) deciding what phenomena to model; (b) deciding how to model the chosen phenomena; and (c) satisfying domain-dependent constraints. Third, it introduces a new class of approximations called *causal approximations*, that are commonly found in modeling the physical world. The basic idea underlying the definition of causal approximations is that more approximate descriptions usually explain less about a phenomenon than more accurate descriptions. As a consequence, the causal relations entailed by a model decrease monotonically as models become simpler. This leads to the development of an efficient, polynomial-time algorithm for finding adequate models when all approximations are causal approximations.

1 Introduction

One of the earliest important ideas in Artificial Intelligence is :.at effective problem solving requires the use of adequate models of the domain [2]. Adequate models incorporate abstractions and approximations that are well suited to the problem solving task. Different types of abstractions and approximations have been identified for a variety of tasks: abstractions in ABSTRIPS speed up planning by dropping select operator preconditions [32]; approximations in mathematical domains simplify equation solving by ignoring negligible quantities [4; 31]; piecewise-linear approximations of ordinary differential equations are used in PLR to analyze dynamic engineering systems [33]; fitting approximations support efficient model sensitivity analysis [42; 43]; horn approximations of a logical theory allow efficient inference [35]. In this paper we introduce a new class of approximations, called *causal approximations*, that are commonly found in modeling the physical world. Causal approximations support the efficient generation of ***parsimonious causal explanations***.

Parsimonious causal explanations play an important role in reasoning about engineered devices [45]. On the one hand, they are a vehicle for explaining phenomena of interest to a human user. On the other hand, they can be used to focus subsequent reasoning: in design, causal explanations allow the identification of changes to be made to a device to create a better design; in diagnosis, causal explanations focus the reasoning on just what could have caused a symptom; causal explanations can focus quantitative analysis.

Causal explanations are usually generated from underlying device models [11; 15; 21; 23; 37; 47]. Hence, to generate parsimonious causal explanations, the underlying device models must be as simple as possible. Device models can introduce irrelevant detail into causal explanations either by modeling irrelevant phenomena, or by including needlessly complex models of relevant phenomena. Consider the temperature gauge in Figure 1, consisting of a battery, a wire, a bimetallic strip, a pointer, and a thermistor. A thermistor is a semiconductor device; an increase in its temperature causes a decrease in its resistance. A bimetallic strip has two strips made of different metals welded together. Temperature changes cause the two strips to expand by different amounts, causing the bimetallic strip to bend. The following is a causal explanation of how the gauge works: the thermistor's temperature determines its resistance. This determines the circuit current, which determines the heat generated in the wire, and hence the bimetallic strip's temperature. This determines the bimetallic strip's deflection, which determines the pointer's angular position.

To generate the above explanation, we model the wire as a resistor that dissipates heat due to current flow. Modeling irrelevant phenomena (e.g., the electromagnetic field generated by the wire) is unnecessary. Approximating the wire's resistance by assuming it is constant is adequate—more accurate models that include the dependence of the wire's resistance on its temperature or length are unnecessary.

No single device model is adequate for generating parsimonious explanations for all phenomena: every model will be either unnecessarily complex or too simple to

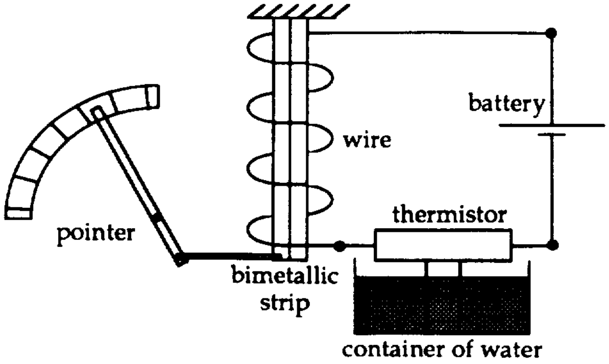

Figure 1: A temperature gauge

explain some phenomenon. For example, suppose we wanted to redesign the above temperature gauge to make its functioning independent of the atmospheric temperature. To successfully achieve this goal, we need the following parsimonious causal explanation for how the atmospheric temperature affects the angular position of the pointer: the atmospheric temperature affects the bimetallic strip's temperature. This determines the deflection of the bimetallic strip, which determines the angular position of the pointer. Note that this explanation used a much simpler model of the temperature gauge than the previous explanation, e.g., the electrical aspects of the device were deemed irrelevant. The resulting parsimony of the explanation allows the redesign effort to focus on just the relevant phenomena.

In this paper we analyze the problem of automatically selecting adequate models for physical systems, with a focus on the task of generating parsimonious causal explanations.[1] This analysis yields three important contributions. First, we develop a precise formalization of the problem of finding models that generate parsimonious causal explanations. Second, we use this formalization to show that, in general, the problem is intractable, while also identifying the sources of intractability. Third, we introduce a new class of approximations called causal approximations, that are commonly found in modeling the physical world. We show that when all approximations are causal approximations, the above problem can be solved efficiently.

We cast the problem of automatically selecting adequate models as a search problem, requiring answers to the following three questions:

- *What is a model, and what is the space of possible models?* (What is the search space?)
- *What is an adequate model?* (What is the goal criterion?)
- *How do we search the space of possible models for adequate models?* (What is the search strategy?)

[1] For the purposes of this paper "device" and "physical system" are assumed to be synonyms.

1.1 What is a model and what is the space of possible models?

Models of physical systems are best expressed as a set of equations that describe various physical phenomena occurring in the system.[2] However, rather than viewing a model as just a set of equations, we will view it as a set of *model fragments* [14]. A model fragment is a set of equations that partially describe a single device phenomenon at some level of detail. The set of such model fragments, (partially) describing different aspects of the device, can be collected together in a library. This library is a compact description of the space of all device models: different subsets of the library correspond to different models, i.e., the model fragments in the library are the "building blocks" out of which models are constructed. Model fragments are the appropriate building blocks because: (a) they are easier to create than complete models; (b) they can be reused in different models; and (c) not all meaningful physical phenomena can be represented by a single equation. Section 2 discusses the details of models and model fragments.

1.2 What is an adequate model?

The task of generating parsimonious causal explanations places two essential requirements on adequate models. First, they must be able to provide a causal explanation for the phenomenon of interest. We construct causal explanations from a model by generating the causal ordering of the parameters of the model using the equations of the model. Second, the causal explanations must be parsimonious, i.e., the model generating the explanations must be as simple as possible. We define the simplicity of a model based on an *approximation* relation between model fragments. This definition is based on the intuition that (a) a model is simpler if it models fewer phenomenon; and (b) approximate descriptions are simpler than more complex ones. In addition to these two fundamental criteria, we also require that adequate models satisfy any user-specified domain-dependent constraints. Section 3 presents a detailed discussion of our definition of model adequacy.

1.3 How do we find adequate models?

Given a library of model fragments, there is an exponentially large space of possible device models. We will show in Section 4 that the problem of finding an adequate model in this space of possible models is intractable (NP-hard). Intuitively, this means that, to find an adequate model, we can do little better than to check a significant fraction of this exponentially large space. Even for small systems, this space is extremely large, so any brute force approach is out of the question. However, this appears to contradict the observation that domain experts are able to provide

[2] Point of terminology: this notion of a model differs from the notion of a model used in mathematical logic. Our notion of a model corresponds roughly to a logical *theory*.

parsimonious causal explanations after only a little bit of thought. This means that the world provides additional structure, which can be exploited to develop an efficient model selection algorithm.

In Section 5 we identify an important source of such additional structure. In particular, we introduce *causal approximations*, a new class of approximations that form the basis of an efficient model selection algorithm. The basic idea underlying our definition of causal approximations is that more approximate descriptions often tend to involve fewer parameters. Furthermore, approximate descriptions tend to explain less about a phenomenon than more accurate descriptions. An important consequence of these properties is that when all approximations are causal approximations, the causal relations entailed by a model decrease monotonically as the model become simpler. Hence, if a model does not explain the phenomenon of interest, neither does any simpler model. This leads to an efficient model selection algorithm, based on simplifying the most accurate model as much as possible. Causal approximations are particularly useful because they are commonly found in modeling the physical world. Appendix A presents a number common approximations, all of which are causal approximations.

The treatment in Section 5 is restricted to models that do not contain differential equations. Section 6 generalizes these results to models that contain differential equations.

2 Models and model fragments

We will be concerned with models of the behavior of physical systems, typically of engineered devices. Models of device behavior are best represented as a set of *equations* that relate a set of *parameters*. In this paper we will only consider *lumped parameter* models, which disregard the dependence of parameter values on spatial location. However, we will consider both time-varying models, represented using ordinary differential equations, and equilibrium models, represented using algebraic and qualitative equations [5; 24].[3]

A device can be modeled in many different ways, i.e., it can be described by different sets of equations. Different models can model different phenomena, or can use different models for the same phenomena. Since a device can be modeled in a variety of different ways, it is important that we be able to represent the space of possible models of the device. We represent this space using *model fragments*.

2.1 Model fragments

A model fragment is a set of independent equations that partially describe some physical phenomena. Different model fragments can describe different phenomena, or can be different descriptions of the same phenomena. For example, Figure 2 shows

[3] Most of the research in qualitative reasoning about physical systems makes these assumptions.

a model fragment that describes electrical conduction in a wire by modeling the wire as a resistor. Figure 3 shows a different model fragment that describes the same phenomenon for the wire by modeling the wire as an ideal conductor. Finally, Figure 4 shows a model fragment that describes the temperature dependence of the wire's length, a completely different phenomenon. (The equation $exogenous(q)$ represents the fact that the value of q is determined exogenously by a mechanism that has not been explicitly modeled; it can be viewed as a shorthand for the equation $q = c$ for some constant c.)

$$\{V_w = i_w R_w\}$$

Figure 2: Model fragment describing a wire as a resistor.

$$\{V_w = 0\}$$

Figure 3: Model fragment describing a wire as an ideal conductor.

$$\begin{gathered}\{l_w = l_{w0}(1 + \alpha_w(T_w - T_{w0})),\\ exogenous(\alpha_w),\\ exogenous(l_{w0})\}\end{gathered}$$

Figure 4: Model fragment describing the temperature dependence of the wire's length.

Model fragments, which can be viewed as either component model instances [11; 47], or process instances [15][4], usually have *applicability* conditions (e.g., operating conditions [11; 47] or quantity conditions [15]). There are well developed techniques for handling such applicability conditions [7; 16; 22]. Hence, for simplicity, we will not explicitly model and reason about these applicability conditions; rather we assume that the only model fragments under consideration are the ones whose applicability conditions are satisfied.

2.2 Models

A device model is constructed by composing a set of model fragments, i.e., rather than viewing a model just as a set of equations, it is much more useful to think of it as a set of model fragments. The equations of a model, viewed as a set of model fragments, is just the union of the equations of the model fragments in the model,

[4] This is in contrast to model fragments in [14] which are class level, rather than instance level, descriptions of phenomena.

i.e., $E(M)$, the equations of a model M, is:

$$E(M) = \bigcup_{m \in M} m \tag{1}$$

Other methods of combining equations are also possible, e.g., influence combination in [15]. Appendix B shows that the results of this paper can be easily extended to handle these methods.

The primary advantage of viewing models as sets of model fragments is that the set of applicable model fragments is an implicit representation of a very large set of device models. This is because any subset of this set of model fragments can be composed to form a model.[5] Hence, a library of model fragments is a compact representation of an exponentially large set of models. Alternate representations of this large space of models, e.g., by explicitly representing each model [1], are unrealistic. In the rest of this paper, we will let $\mathcal{M}$ denote the set (or library) of applicable model fragments, with every device model being a subset of $\mathcal{M}$.

Another important advantage of model fragments is that, as compared to individual equations, they are better suited to be the "building blocks" of models. This is because model fragments allow us to collect together related sets of equations into a single unit, making it possible to easily represent phenomena that cannot conveniently be represented by a single equation.

2.3 Relations between model fragments

The model fragments in $\mathcal{M}$ are related to each other with the *contradictory* and *approximation* relations, and are organized into *assumption classes*. These concepts will prove to be important in the next section, where we give a precise definition of model adequacy.

2.3.1 The *contradictory* relation

As mentioned earlier, different model fragments can be descriptions of different phenomena, or can be different descriptions of the same phenomena. When model fragments describe the same phenomena, they often make contradictory assumptions about the domain. For example Figure 5 shows three different model fragments describing electrical conduction in `wire-1`, which make mutually contradictory assumptions. In particular, the first assumes that the resistance of the conductor is zero, the second assumes that the resistance of the conductor is infinite, while the third assumes that the resistance of the conductor is non-zero and finite.

We represent the fact that model fragments make contradictory assumptions about the domain using the *contradictory* relation. If m_1 and m_2 are model fragments, then *contradictory*(m_1, m_2) says that m_1 and m_2 make contradictory assumptions about

[5] As we shall see later, not every subset of model fragments is an adequate model, but the basic observation still holds.

`Ideal-conductor(wire-1):` $\{V_w = 0\}$
`Ideal-insulator(wire-1):` $\{i_w = 0\}$
`Resistor(wire-1):` $\{V_w = i_w R_w\}$

Figure 5: Model fragments describing electrical conduction in `wire-1`.

the domain. It is important to note that the contradiction between *contradictory* model fragments cannot, in general, be derived from the equations of the model fragments. For example, there is nothing intrinsically contradictory about the equations of the first and second model fragments above, i.e., it is certainly possible that both the current through a conductor and the voltage drop across the conductor is zero. The contradiction between these model fragments is a consequence of a contradiction between the assumptions underlying them. The *contradictory* relation is a simple way of representing the contradiction between the underlying assumptions.

Clearly the *contradictory* relation is irreflexive (so that model fragments cannot contradict themselves), and symmetric (so that model fragments are mutually contradictory):

$$\neg \mathit{contradictory}(m_1, m_1) \tag{2}$$

$$\mathit{contradictory}(m_1, m_2) \Rightarrow \mathit{contradictory}(m_2, m_1) \tag{3}$$

2.3.2 The *approximation* relation

In addition to specifying that model fragments contradict each other, a domain expert may be able to specify that one model fragment is a more *approximate* description of a phenomenon than another. This means that the predictions made by the more accurate model fragment are "closer to reality" than the predictions made by the more approximate model fragment. We represent such knowledge using the *approximation* relation between model fragments: $\mathit{approximation}(m_1, m_2)$ says that the model fragment m_2 is a more approximate description of some phenomena than the model fragment m_1. For example, Figure 6 shows some of the approximation relations between the model fragments shown in Figure 5.

approximation(`Resistor(wire-1)`, `Ideal-conductor(wire-1)`)
approximation(`Resistor(wire-1)`, `Ideal-insulator(wire-1)`)

Figure 6: Approximation relation between the electrical conduction model fragments.

Once again, it is important to note that the *approximation* relation is a primitive, domain-dependent relation, and this relation cannot be derived directly from the equations of the model fragments. For example, there is nothing about the equations of the ideal conductor model fragment that tells us that it is necessarily a more approximate description of electrical conduction than the resistor model fragment; this just happens to be a domain fact discovered by scientists and engineers.

Clearly the *approximation* relation is irreflexive, anti-symmetric, and transitive (so that model fragments are not approximations of themselves, and *approximation* forms a partial ordering on the relative accuracy of the model fragments describing a phenomena):

$$\neg \mathit{approximation}(m_1, m_1) \tag{4}$$

$$\mathit{approximation}(m_1, m_2) \Rightarrow \neg \mathit{approximation}(m_2, m_1) \tag{5}$$

$$\mathit{approximation}(m_1, m_2) \wedge \ \mathit{approximation}(m_2, m_3) \Rightarrow \mathit{approximation}(m_1, m_3) \tag{6}$$

Furthermore, since approximations make different, and hence contradictory, predictions about the same phenomenon, we will require that all approximations are also mutually contradictory:

$$\mathit{approximation}(m_1, m_2) \Rightarrow \mathit{contradictory}(m_1, m_2) \tag{7}$$

2.3.3 Assumption classes

An *assumption class* is a set of mutually contradictory model fragments, i.e., if m_1 and m_2 are model fragments, and A is an assumption class, we have:

$$(m_1, m_2 \in A) \ \wedge \ m_1 \neq m_2 \Rightarrow \mathit{contradictory}(m_1, m_2) \tag{8}$$

An assumption class can be viewed as a modeling dimension. This view highlights the fact that, to avoid using mutually *contradictory* model fragments, a choice needs to be made along the modeling dimension represented by the assumption class. One can see that the model fragments in Figure 5 form an assumption class describing electrical conduction in the wire.

3 Adequate models

The adequacy of a model is closely tied to the task for which the model is to be used. Simulations carried out during the final stages of detailed design require the use of high fidelity models that incorporate accurate, quantitative descriptions of all significant phenomena. On the other hand, models that support analysis during conceptual design can be much coarser. Similarly, Hamscher [19, page 11] argues that:

> For complex devices the model of the target device should be constructed with the goal of troubleshooting explicitly in mind.

In this paper we define the adequacy of a model with respect to the task of generating parsimonious causal explanations for phenomena of interest. Causal explanations play an important role in automated reasoning systems as a vehicle for the system to communicate with its human users. Such explanations can be used for instructional purposes, as in various Intelligent Computer Aided Instruction systems [6; 17;

41], or as a method for explaining the system's line of reasoning to a human user [30; 39; 40].

In addition to their role in communication, causal explanations play a central role in focusing other forms of reasoning [45]. Causal explanations are used in diagnosis to focus the reasoning only on those elements that could have caused a particular symptom [9]. Causal explanations focus design and redesign by focusing the reasoning on just those mechanisms that can produce the desired behavior [48]. Causal explanations can also guide quantitative analysis by providing an overall structure for solving the problem at hand [10].

3.1 Causal explanations as causal ordering

Different types of causal explanations are generated depending on the particular vocabulary used for modeling the causal relation. In this paper we adopt the vocabulary commonly used in the literature on qualitative reasoning about physical systems [45]: the causal relation relates *parameters*, and the causal relation represents a dependence of the value of the "effect" parameter on the "cause" parameter. This causal dependence between parameters induces a partial ordering on the parameters, called a *causal ordering*.

The causal dependence between parameters can take one of two forms: *functional dependency* and *integration*. The functional dependency of a parameter p_1 on a parameter p_2 corresponds to a causal mechanism that "instantaneously" determines the value of p_1 as a function of the value of p_2 (and, possibly, some other parameters). We have quoted the "instantaneously" to emphasize that what counts as "instantaneous" is a modeling decision related to the time scale of interest [21; 25]. Causal relations as functional dependencies were first studied in [37], and subsequently in [11; 23; 47] and in [15], where they are called *indirect influences*.

The other type of causal relations between parameters is the integration relation between a parameter and its derivative. In contrast to functional dependencies that act instantaneously, the integration relation acts over a period of time. Causal relations as integration have been studied in [21] and in [15], where they are called *direct influences*.

3.2 Constructing the causal ordering

The causal ordering of a set of parameters is generated from the equations of a device model. Functional dependencies are generated from algebraic and qualitative equations, while integration relations are generated from differential equations. In this and the next two sections we will discuss only the former; differential equations will be discussed in Section 6.

Equations, as such, can be viewed as acausal representations of domain mechanisms. For example, the equation $V = iR$ (Ohm's law) is an acausal representation of a mechanism for electrical conduction. It merely states that the voltage across an electrical conductor, V. is proportional to the current through the conductor, i, with the resistance of the conductor, R, being the proportionality constant. However, it makes no causal claims like "the voltage depends on the current."

3.2.1 Causal orientations of equations

To have a causal import, equations must be causally oriented. A causally oriented equation represents the fact that one of the parameters of the equation is directly causally dependent on the other parameters of the equation. The dependent parameter is said to be causally determined by the equation. For example, the acausal equation $V = iR$ can be causally oriented so that it causally determines V, making V directly causally dependent on i and R.

The causal orientation of an equation can be fixed *a priori* [15], or it can be inferred from the equations comprising a model of the system [11; 21; 23; 37; 47]. Fixing the causal orientation of each equation *a priori* is overly restrictive, since different causal orientations are often possible. However, not all causal orientations fit a domain experts intuitions about causality. For example, the equation $V = iR$ can be causally oriented in one of two ways: either V can be causally dependent on i and R, or i can be causally dependent on V and R. However, the third possibility, R being causally dependent on V and i, makes no sense because, in an ordinary electrical conductor, there is no way that changing V and/or i can cause a change in R.

The set of allowed causal orientations of an equation, e, can be represented by the set, $P_c(e)$, of parameters that can be causally determined by e. As a typographical aid, parameters that can be causally determined by an equation will be typeset in boldface, e.g., $\boldsymbol{V} = \boldsymbol{i}R$ says that this equation can causally determine V and i but not R. We extend the function P_c to a set E of equations, and to a model M, in the natural way (recall that a model fragment is just a set of equations):

$$P_c(E) = \bigcup_{e \in E} P_c(e) \tag{9}$$

$$P_c(M) = \bigcup_{m \in M} P_c(m) \tag{10}$$

In addition, let $P(e)$ be the set of all parameters in equation e. As with P_c, extend P to a set E of parameters, and to a model M, as follows:

$$P(E) = \bigcup_{e \in E} P(e) \tag{11}$$

$$P(M) = \bigcup_{m \in M} P(m) \tag{12}$$

3.2.2 Causal mappings

Serrano and Gossard [36] make the key observation that, given a set of equations, the causal ordering of the parameters can be generated efficiently by (a) causally orienting each equation such that each parameter is causally determined by exactly one equation; and (b) taking the transitive closure of the direct causal dependencies entailed by the causal orientations.[6]

We formalize Serrano and Gossard's observation by first defining a *causal mapping*:

Definition 1 (Causal mapping) *Let E be a set of equations. A function $F : E \rightarrow P(E)$ is said to be a causal mapping if and only if (a) F is 1-1; and (b) for each $e \in E$, $F(e) \in P_c(e)$. F is an onto causal mapping if for each parameter $p \in P(E)$, there is an equation $e \in E$, such that $F(e) = p$.*

Hence, a causal mapping causally orients each equation such that each parameter is causally determined by at most one equation, while an onto causal mapping causally determines every parameter.

Note that the co-domain of F in the above definition is $P(E)$ and not $P_c(E)$, even though condition (b) guarantees that the range of F is a subset of $P_c(E)$. We have chosen $P(E)$ as the co-domain of F to ensure that when F is onto then each parameter in $P(E)$ is causally determined by an equation in E.

The direct causal dependencies entailed by a causal mapping $F : E \rightarrow P(E)$ is denoted by C_F, and is defined as follows:

$$C_F = \{(p_1, p_2) : (\exists e \in E)\ F(e) = p_2\ \wedge\ p_1 \in P(e)\} \tag{13}$$

In other words, $(p_1, p_2) \in C_F$ if and only if p_2 directly causally depends on p_1 in the causal orientations defined by F. Denote the transitive closure of C_F by $tc(C_F)$. The following lemma states that the transitive closure of different onto causal mappings of E are identical. (We will soon discuss conditions under which onto causal mappings exist.)

Lemma 1 *Let E be a set of independent equations, and let $F_1 : E \rightarrow P(E)$ and $F_2 : E \rightarrow P(E)$ be onto causal mappings. Then $tc(C_{F_1}) = tc(C_{F_2})$.*

Proof: It suffices to show that $C_{F_1} \subseteq tc(C_{F_2})$. Let $(q, p) \in C_{F_1}$, and let $e \in E$ such that $F_1(e) = p$, and hence $q \in P(e)$. We show that $(q, p) \in tc(C_{F_2})$. Construct the sequence $p_0, p_1, \ldots, p_m$ such that (a) $p_0 = p$; (b) $p_i = F_2(F_1^{-1}(p_{i-1}))$, for $1 \leq i \leq m$; (c) p_m is the first repetition in the sequence, i.e., $p_i \neq p_j, 0 \leq i, j \leq (m-1), i \neq j$, and $p_m = p_i$, for some $i, 0 \leq i \leq (m-1)$. Such a sequence must exist because F_1 and F_2 are onto causal mappings, and because there are a finite number of parameters. We claim that $p_m = p_0$, for if $p_m = p_i$ for some i, $1 \leq i \leq (m-1)$, then $p_{m-1} = F_1(F_2^{-1}(p_m)) = F_1(F_2^{-1}(p_i)) = p_{i-1}$, which contradicts condition (c) above.

[6] Serrano and Gossard do not actually talk about causal ordering or causal orientations. They are interested in efficiently evaluating a set of constraints. However, the parameter dependencies that they generate are identical to the causal ordering, and their algorithm can be viewed as causally orienting each equation. Hence, we attribute the above observation to them.

Next, let $e_i = F_1^{-1}(p_{i-1})$, for $1 \leq i \leq m$. Hence, $p_{i-1} \in P(e_i)$ and $p_i = F_2(e_i)$, so that $(p_{i-1}, p_i) \in C_{F_2}$. Hence, by transitivity, $(p_1, p_m) \in tc(C_{F_2})$, and since $p_m = p_0 = p$, we have $(p_1, p) \in tc(C_{F_2})$. Now there are two cases: (a) if $p_1 = q$, then $(q, p) \in tc(C_{F_2})$; or (b) if $p_1 \neq q$, then since $p_1 = F_2(e)$ and $q \in P(e)$, we have $(q, p_1) \in C_{F_2}$, and by transitivity $(q, p) \in tc(C_{F_2})$. □

Intuitively, the above proof shows that if F_1 and F_2 differ on the parameter to which an equation e is mapped, then the parameters $F_1(e)$ and $F_2(e)$ are causally dependent on each other.

3.2.3 Causal ordering

The causal ordering generated from a set of equations is the transitive closure of the direct causal dependencies generated by *any* onto causal mapping of the set of equations:

Definition 2 (Causal order) *Let E be a set of independent equations, and let $F : E \longrightarrow P(E)$ be an onto causal mapping. The causal order of the parameters of E, denoted $C(E)$, is the transitive closure of C_F:*

$$C(E) = tc(C_F)$$

The causal ordering is well defined because Lemma 1 assures us that the transitive closures of all onto causal mappings of a set of equations are identical.

Figure 7 shows a set of equations describing the temperature gauge shown in Figure 1. Figure 8 shows an onto causal mapping, while Figure 9 shows a graphical representation of the direct causal dependencies generated from this onto causal mapping. The transitive closure of these direct causal dependencies corresponds to the causal order generated by the equations in Figure 7.

Since generating the causal ordering requires an onto causal mapping, a natural question that arises is: under what conditions does an onto causal mapping exist? To answer this question we introduce *complete* sets of equations. Informally, a set of independent equations is complete if it has as many equations as parameters, and no subset of equations has fewer parameters than equations. One can see that if some subset of equations had fewer parameters than equations, then that set is *overconstrained.*[7] More precisely, we have the following definitions:

Definition 3 *Let E be a set of independent equations. E is said to be complete if and only if (a) $|E| = |P_c(E)| = |P(E)|$; and (b) for every $S \subseteq E, |S| \leq |P_c(S)|$.*[8] *$E$ is said to be overconstrained if and only if there exists $S \subseteq E$ such that $|S| > |P_c(S)|$.*

[7] Being independent, the possibility of the equations being merely redundant is ruled out.

[8] "$|\cdot|$" returns the cardinality of a set.

`Linkage(bms-1,ptr-1):`	$\boldsymbol{\theta}_p = k_1 \boldsymbol{x}_b$
`Thermal-bms(bms-1):`	$\boldsymbol{x}_b = k_2 T_b$
`Heat-flow(bms-1,atm-1):`	$\boldsymbol{f}_{ba} = k_3(\boldsymbol{T}_b - \boldsymbol{T}_a)$
`Heat-flow(wire-1,bms-1):`	$\boldsymbol{f}_{wb} = k_4(\boldsymbol{T}_w - \boldsymbol{T}_b)$
`Constant-temperature(atm-1):`	$exogenous(\boldsymbol{T}_a)$
`Thermal-equilibrium(bms-1):`	$\boldsymbol{f}_{ba} = \boldsymbol{f}_{wb}$
`Thermal-equilibrium(wire-1):`	$\boldsymbol{f}_{wb} = \boldsymbol{f}_w$
`Resistor(wire-1):`	$\boldsymbol{V}_w = \boldsymbol{i}_w \boldsymbol{R}_w$
`Constant-resistance(wire-1):`	$exogenous(\boldsymbol{R}_w)$
`Thermal-resistance(wire-1):`	$\boldsymbol{f}_w = V_w i_w$
`Electrical-thermistor(thermistor-1):`	$\boldsymbol{V}_t = \boldsymbol{i}_t R_t; \quad \boldsymbol{R}_t = k_5 e^{k_6/T_t}$
`Constant-voltage-source(battery-1):`	$exogenous(\boldsymbol{V}_v)$
`Kirchhoff's laws:`	$\boldsymbol{V}_v = \boldsymbol{V}_w + \boldsymbol{V}_t; \quad \boldsymbol{i}_v = \boldsymbol{i}_t; \quad \boldsymbol{i}_t = \boldsymbol{i}_w$
`Input:`	$exogenous(\boldsymbol{T}_t)$

θ_p: Pointer angle	x_b: Bms deflection	R_w: Wire resistance
R_t: Thermistor resistance	i_t: Thermistor current	V_t: Thermistor voltage
i_w: Wire current	V_w: Wire voltage	i_v: Battery current
V_v: Battery voltage	T_b: Bms temperature	T_a: Atm temperature
T_w: Wire temperature	T_t: Thermistor temperature	f_{ba}: Heat flow (bms to atm)
f_{wb}: Heat flow (wire to bms)	f_w: Heat generated in wire	k_j: Exogenous constants

Figure 7: A possible model of the temperature gauge

The following lemma states that an onto causal mapping exists if and only if the set of equations is complete. This means that when a set of equations is complete, all the parameters in the equations can be causally determined.

Lemma 2 *Let E be a set of independent equations. Then there exists an onto causal mapping $F : E \to P(E)$ if and only if E is complete.*

Proof: Construct the bipartite graph $G = (X, Y, R)$[9] with $X = E$, $Y = P(E)$, and an edge in R connects a node (equation) $e \in X$ to a node (parameter) $p \in Y$ if and only if $p \in P_c(e)$.[10] A *matching* in G is a subset of R such that no two edges in the subset share a common node. A *complete* matching is a matching in which there is an edge incident on every node. It is easy to see that an onto causal mapping on E corresponds exactly to a complete matching in G. Hall's Theorem [12, pages 137-138] tells us that G contains a complete matching if and only if (a) $|X| = |Y|$; and (b) for every $A \subseteq X, |A| \leq |R(A)|$, where $R(A)$ denotes the set of nodes connected to the nodes in A by edges in R. This is equivalent to stating that E is a complete set of equations. □

[9] $X \cup Y$ is the set of nodes and R is the set of edges. Each edge in R connects a node in X to a node in Y.

[10] This bipartite graph representation of the set of equations is due to [36].

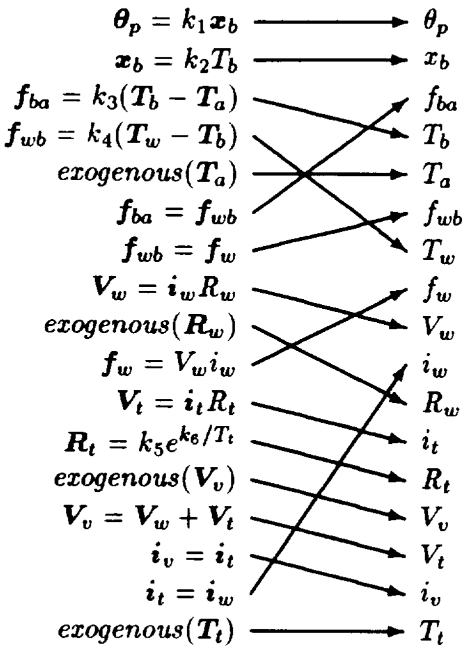

Figure 8: An onto causal mapping.

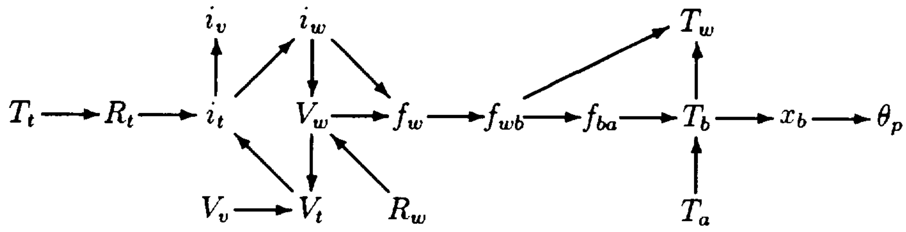

Figure 9: The direct causal dependencies generated by the above onto causal mapping

The above proof also suggests an efficient algorithm for constructing onto causal mappings: construct a complete matching in the bipartite graph G defined above. An efficient (polynomial time) algorithm for finding complete matchings in bipartite graphs is described in [12]. Constructing the causal ordering from the causal mapping is, of course, easy.

Having discussed causal ordering and how they are constructed from a device model, we are now in a position to precisely define model adequacy. In the next four subsections we will discuss the consistency and completeness of models, the expected behavior. ·main dependent constraints, and model simplicity. We will conclude with a precise statement of the problem of finding an adequate model.

3.3 Consistency and completeness of models

Any consistent device model must make a consistent set of assumptions about the domain. Recall that when two model fragments make contradictory domain assumptions, they are related by the *contradictory* relation. Therefore, consistent models do not use mutually *contradictory* model fragments. Similarly, in Definition 3 we defined the notion of an overconstrained set of equations. If a set of independent equations is overconstrained, then the equations have no solution, leading to a contradiction. Furthermore, we would like an adequate model to causally determine all the parameters in the model, i.e., we would like the equations of the model to be complete. These observations lead directly to our definition of a consistent and complete model:

Definition 4 (Consistency and completeness) *A model M is said to be consistent if and only if the following two conditions are satisfied:*

1. $\forall m_1, m_2 \in M\ \neg contradictory(m_1, m_2)$*, i.e., the model does not contain mutually contradictory model fragments;*
2. *The set of equations of M is not overconstrained.*

A model M is said to be complete if and only if the set of equations of M is complete.

To ensure that adequate models make a consistent set of domain assumptions, while providing complete descriptions of all phenomena, we have:

- An adequate model must be consistent and complete.

If we assume that the model in Figure 7 contains no mutually *contradictory* model fragments, then it is both consistent and complete because the set of equations of the model is complete (and hence not overconstrained).

3.4 Representing the phenomenon of interest

As discussed earlier, we will be defining model adequacy with respect to the task of providing parsimonious causal explanations for the phenomenon of interest. Hence, the phenomenon of interest is a crucial input that focuses model selection. We call the phenomenon of interest the *expected behavior*. The expected behavior of a device is an abstract description of *what* the device does (but not *how* it does it). The causal explanation generated by a model is a description of how the expected behavior is achieved.

Following our discussion of causal ordering, we specify expected behaviors as a query that requests a causal explanation for how one parameter causally depends on another. For example, the expected behavior of the temperature gauge shown in Figure 1, representing its primary function, is:

$$causes(T_t, \theta_p)$$

This expected behavior requests a causal explanation for how the thermistor's temperature causally determines the pointer's angular position.

The expected behavior provides us with our most important criterion for model adequacy:

- An adequate model must explain the expected behavior, i.e., a model is adequate with respect to an expected behavior, $causes(p_1, p_2)$, if it is able to provide a causal explanation for how p_2 causally depends on p_1.

Since we can efficiently generate the causal ordering from a set of equations, it is easy to check whether or not a model satisfies the expected behavior. The model in Figure 7 is adequate with respect to the expected behavior $causes(T_t, \theta_p)$ because Figure 9 shows that θ_p is dependent on T_t in the corresponding causal ordering.

It must be noted that our language for expressing the expected behaviors is extremely simple; it only allows us to ask for explanations for causal dependencies between parameters. More expressive languages are, of course, desirable. We might want to include additional qualitative information, e.g., increasing T_t causes θ_p to increase. Or we might want to include more information about the actual functional relationship, e.g., a linear relationship between T_t and θ_p.

However, the price we must pay for using more expressive languages for the expected behavior is that checking whether or not the expected behavior is satisfied becomes very expensive, and can often even be impossible. For example, deciding whether an increase in T_t causes an increase or a decrease in θ_p with purely qualitative information is not possible when there are competing influences. Additional information about the relative magnitudes of these influences is necessary, which may or may not be available. Hence, we have chosen a simple, though useful, language for expressing the expected behavior, leading to an efficient algorithm for deciding whether or not a model satisfies the expected behavior.

3.5 Domain-dependent constraints

A domain expert often requires that an adequate model must satisfy a set of domain-dependent constraints. For example, recall that model fragments are partial descriptions of phenomena. Additional model fragments are required to complete this description. A domain expert may require that the partial description of a model fragment must be completed by the use of model fragments in a specified assumption class, e.g., the partial description of electrical conduction specified by the model fragment **Resistor(wire-1)** in Figure 5 must be completed by the inclusion of a model fragment from an assumption class describing the wire's resistance. (See [14; 28; 29] for other examples).

We express such domain-dependent constraints using *propositional coherence constraints*. A propositional coherence constraint is just a propositional formula in which the propositions are model fragments. A propositional coherence constraint is satisfied with respect to a model M just in case the corresponding propositional formula is satisfied by the interpretation that assigns *true* to a proposition if and only if the proposition is in M, and *false* otherwise. For example, the propositional coherence constraint

$$(m_1 \vee m_2) \Rightarrow m_3$$

is satisfied by the model $\{m_1, m_3\}$, the model $\{m_2, m_3\}$, the empty model, etc.

As a convenient shorthand, we allow the use of assumption classes in propositional coherence constraints. Recall that an assumption class is a set of mutually contradictory model fragments. Hence, we use an assumption class as a shorthand for a disjunction of the model fragments in the assumption class. For example, if the assumption class A contains the model fragments m_4 and m_5, then the propositional coherence constraint

$$m_3 \Rightarrow A$$

is equivalent to the propositional coherence constraint

$$m_3 \Rightarrow (m_4 \vee m_5)$$

Let $\mathcal{C}$ denote the set of all domain-dependent constraints. We have the following:

- An adequate model must satisfy every propositional coherence constraint in the set $\mathcal{C}$ of all domain-dependent constraints.

3.6 Simplicity of models

Thus far, we have said that an adequate model must be consistent and complete, must be able to explain the expected behavior, and must satisfy all domain-dependent constraints. Typically a very large number of device models satisfy these criteria. Most of these models introduce irrelevant detail into the causal explanations they generate, either by modeling irrelevant phenomena, or by including needlessly complex models of relevant phenomena.

For example, assume that the model in Figure 7 satisfies all the above criteria. Other models that augment this model by modeling additional phenomena, such as the electromagnetic field generated by the wire, would also satisfy the above criteria. Similarly, models that use more accurate descriptions of phenomena that are already modeled, e.g., by modeling the wire as a temperature dependent resistor rather than a constant resistance resistor, would also satisfy the above criteria. Such models introduce irrelevant detail into the causal explanation of how the thermistor's temperature affects the pointer's angular position.

To address this problem we need a ***simplicity ordering*** on the models. Given such a simplicity ordering, we will say that an adequate model is a simplest model that satisfies all the above criteria, i.e., no simpler model satisfies the above criteria. The simplicity ordering we consider is a partial ordering of the models, and is based on the *approximation* relation between model fragments. This definition of simplicity is based on the following two intuitions: (a) a model is simpler if it models fewer phenomena; and (b) approximate descriptions are simpler than more accurate ones.

Definition 5 (Simplicity of models) *A model M_2 is simpler than a model M_1 (written $M_2 \leq M_1$) if for each model fragment $m_2 \in M_2$ either (a) $m_2 \in M_1$; or (b) there is a model fragment $m_1 \in M_1$ such that m_2 is an approximation of m_1, i.e., approximation(m_1, m_2). M_2 is strictly simpler than M_1 (written $M_2 < M_1$) if $M_2 \leq M_1$ and $M_1 \not\leq M_2$.*

It is important to note that this definition of model simplicity is based purely on the intuitions mentioned above. In particular, the definition does not guarantee that a simpler model is more efficient. Nor does it guarantee that simpler models lead to simpler causal explanations of the expected behavior. However, while there are no such guarantees, we believe that the above definition of simplicity provides a good heuristic for identifying more efficient models, and for generating simpler causal explanations. In particular, it is common engineering practice to simplify models by disregarding irrelevant phenomena and by using all applicable approximations. Furthermore, in Section 5 we introduce causal approximations, which will ensure that the the above definition of simplicity will, in fact, lead to simpler causal explanations.

We will require that adequate models be as simple as possible:

- An adequate model is a simplest model that meets all the criteria discussed above.

3.7 Problem statement

We can now provide a precise statement of the problem of finding an adequate model. The input to this problem is a tuple $\mathcal{I}$:

$$\mathcal{I} = (\mathcal{M}, \mathit{contradictory}, \mathit{approximation}, \mathcal{A}, \mathcal{C}, p, q) \tag{14}$$

where $\mathcal{M}$ is the set of all applicable model fragments, *contradictory* and *approximation* are binary relations on model fragments that satisfy Equations 2–7, $\mathcal{A}$ is the set of all assumption classes, $\mathcal{C}$ is the set of domain-dependent propositional coherence constraints, and p and q are parameters such that *causes*(p, q) is the expected behavior. Using the elements of $\mathcal{I}$, we define coherent, causal, and adequate models. A *coherent* model is a consistent and complete model which satisfies all domain dependent constraints.

Definition 6 (Coherent models) *A model $M \subseteq \mathcal{M}$ is said to be a coherent model if and only if the following conditions are satisfied:*

1. *M contains no mutually contradictory model fragments.*

2. *The equations of M are complete (and hence not overconstrained).*

3. *All the constraints in $\mathcal{C}$ are satisfied by M.*

A *causal* model is a coherent model that also explains the expected behavior.

Definition 7 (Causal model) *A model $M \subseteq \mathcal{M}$ is a causal model, with respect to the expected behavior causes(p, q), if and only if (a) M is a coherent model; and (b) q causally depends on p in the causal ordering generated from the equations of M, i.e., $(p, q) \in C(E(M))$.*

Finally, an *adequate* model is just a minimal causal model.

Definition 8 (Adequate model) *A model $M \subseteq \mathcal{M}$ is an adequate model if and only if M is a causal model and no coherent model strictly simpler than M is a causal model, i.e., for all coherent models M', such that $M' < M$, M' is not a causal model.*

The above definitions lead to the following statement of the problem of finding an adequate model. We call this problem the MINIMAL CAUSAL MODEL problem:

Definition 9 (MINIMAL CAUSAL MODEL problem) *Let $\mathcal{I}$ be as in Equation 14. Find an adequate model with respect to $\mathcal{I}$.*

To help in analyzing the complexity of the MINIMAL CAUSAL MODEL problem, we introduce the CAUSAL MODEL problem, which is the decision problem corresponding to the MINIMAL CAUSAL MODEL problem. The CAUSAL MODEL problem asks whether or not there exists a causal model, without requiring this causal model to be minimal.

Definition 10 (CAUSAL MODEL problem) *Let $\mathcal{I}$ be as in Equation 14. Does there exist a causal model with respect to $\mathcal{I}$?*

4 Complexity of model selection

In this section we analyze the complexity of the CAUSAL MODEL problem and the MINIMAL CAUSAL MODEL problem. We will show that the CAUSAL MODEL problem is NP-complete and, as an immediate corollary, that the MINIMAL CAUSAL MODEL problem is NP-hard. Since it is strongly believed that P $\neq$ NP, these results imply that the general problem of finding adequate device models is intractable, i.e., there is no polynomial time algorithm for finding adequate device models.

We prove that the CAUSAL MODEL problem is NP-complete by first showing that it is in NP, and then showing that it is NP-hard.

Lemma 3 *The* CAUSAL MODEL *problem is in NP.*

Proof: It is easy (i.e., in polynomial time) to check whether a model contains mutually *contradictory* model fragments, and whether it satisfies all the constraints in $\mathcal{C}$. Using the comments following Lemma 2, it is also easy to check whether the model is complete and whether it satisfies the expected behavior. □

We show that the CAUSAL MODEL problem is NP-hard by showing that three of its special cases are NP-hard. The three special cases will identify three sources for the intractability of the CAUSAL MODEL problem. Informally, the three sources are: (a) deciding what phenomena to model; (b) deciding how to model the chosen phenomena; and (c) ensuring that causal models satisfy all domain-dependent constraints. In the next section, we will use this knowledge to design special cases of the MINIMAL CAUSAL MODEL problem that can be solved in polynomial time.

In each of the three special cases, the *contradictory* relation is restricted to partition the set of model fragments into the set of assumption classes, i.e., two model fragments are in the same assumption class *if* and only if they are mutually *contradictory*, i.e., model fragments in different assumption classes cannot be mutually *contradictory*:

$$(\forall m_1, m_2 \in \mathcal{M})\ m_1 \neq m_2 \Rightarrow (\mathit{contradictory}(m_1, m_2) \equiv (\exists A \in \mathcal{A})\ m_1, m_2 \in A) \quad (15)$$

Hence, we can view the problem of finding a causal model as one involving the following two steps: (a) selecting a set of assumption classes; and (b) selecting a single model fragment from each selected assumption class. Intuitively, this corresponds to deciding which phenomena to model (step (a)), and then deciding how to model the chosen phenomena (step (b)).

The first special case consists of those instances of the CAUSAL MODEL problem that satisfy the following two (additional) conditions: (a) the instance has no propositional coherence constraints; and (b) every causal model of the instance includes a model fragment from each assumption class. Hence, this special case allows us to identify the first source of intractability: choosing a model fragment from each assumption class in a set of selected assumption classes is intractable. More abstractly, even if we knew exactly which phenomena we wanted to model, deciding how to model the chosen phenomena is intractable.

Lemma 4 *The* CAUSAL MODEL *problem is NP-hard even if its instances are required to satisfy the following conditions: (a) Equation 15; (b) $C = \emptyset$; and (c) every causal model of the instance includes a model fragment from each assumption class, i.e., if $M \subseteq \mathcal{M}$ is a causal model and $A \in \mathcal{A}$, then $M \cap A \neq \emptyset$.*

Proof: The proof is based on a reduction from the NP-complete problem ONE-IN-THREE 3SAT [34], a variation of the common 3SAT problem in which an acceptable truth assignment must satisfy exactly one literal in each clause. Let (U, C) be an arbitrary instance of the ONE-IN-THREE 3SAT problem, where $U = \{u_1, \ldots, u_n\}$ is the set of boolean variables, and $C = \{c_1, \ldots, c_m\}$ is the set of three literal clauses. Construct an instance of the CAUSAL MODEL problem as follows.

Introduce a model fragment m_l for each literal l, and a model fragment m:

$$\mathcal{M} = \{m_{u_i} : 1 \leq i \leq n\} \cup \{m_{\bar{u}_i} : 1 \leq i \leq n\} \cup \{m\}$$

Let m_l and $m_{\bar{l}}$ be *contradictory*:

$$\mathit{contradictory}(m_{u_i}, m_{\bar{u}_i}),\ \text{ for } 1 \leq i \leq n$$

To satisfy Equation 15, let $\mathcal{A}$ be defined as follows:

$$\mathcal{A} = \{\{m_{u_i}, m_{\bar{u}_i}\} : 1 \leq i \leq n\} \cup \{\{m\}\}$$

Let *approximation* be the empty relation and let $\mathcal{C} = \emptyset$. Introduce the set $\mathcal{P}$ of $(m + n + 3)$ parameters:

$$\mathcal{P} = \{p_0, p_1, \ldots, p_{m+n+2}\}$$

and let $p = p_0$ and $q = p_{m+n+2}$. Introduce the set E of $(3m + 2n + 3)$ equations:

$$E = (\bigcup_{1 \leq j \leq m} E_j) \cup (\bigcup_{1 \leq i \leq n} F_i) \cup G$$

where E_j contains an equation for each literal in clause c_j, F_i contains an equation for literals u_i and $\bar{u}_i$, and G contains three equations, as follows:

$$\begin{array}{rcl} E_j & = & \{e_{jl} : l \text{ is a literal in clause } c_j\} \\ F_i & = & \{f_{u_i}, f_{\bar{u}_i}\} \\ G & = & \{g_1, g_2, g_3\} \end{array}$$

The parameters of the equations in E are defined as follows:

$$\begin{array}{lll} \text{If } e \in E_j, 1 \leq j \leq m, & P(e) = \{p_j, p_{j+1}\} & P_c(e) = \{p_{j+1}\} \\ \text{If } e \in F_i, 1 \leq i \leq n, & P(e) = \{p_{m+i}, p_{m+i+1}\} & P_c(e) = \{p_{m+i+1}\} \\ & P(g_1) = \{p_0\} & P_c(g_1) = \{p_0\} \\ & P(g_2) = \{p_0, p_1\} & P_c(g_2) = \{p_1\} \\ & P(g_3) = \{p_{m+n+1}, p_{m+n+2}\} & P_c(g_2) = \{p_{m+n+2}\} \end{array}$$

The equations in the model fragments of $\mathcal{M}$ are defined as follows:

$$\begin{array}{rcl} E(m_l) & = & \{e_{jl} : \text{literal } l \text{ is in clause } c_j\} \cup \{f_l\} \\ E(m) & = & \{g_1, g_2, g_3\} \end{array}$$

That completes the reduction. Since $p = p_0$, $q = p_{m+n+2}$, and all equations (except g_1) relate consecutively numbered parameters, it is easy to verify that, in any causal model, the dependence of q on p is mediated by the rest of the parameters:

$$(p =)p_0 \rightarrow p_1 \rightarrow \cdots \rightarrow p_{m+n+1} \rightarrow p_{m+n+2}(= q)$$

Furthermore, the definition of P_c implies that p_{j+1} can be determined only by an equation in E_j, p_{m+i+1} by an equation in F_i, and p_0, p_1, p_{m+n+2} by g_0, g_1, g_2, respectively.[11] Hence, one can see that a model is a causal model if and only if it includes exactly one equation from E_j, one from F_i, and all the equations in G. Since the equations in F_i are only in model fragments m_{u_i} and $m_{\bar{u}_i}$, any causal model must include a model fragment from each assumption class (as required in the lemma).

Consider the following correspondence between truth assignments and models that contain model fragment m: a literal l is *true* if and only if model fragment m_l is in the model. We now show that a truth assignment is acceptable if and only the corresponding model is a causal model. Let $l_{j_1}, l_{j_2}, l_{j_3}$ be the three literals in clause c_j. A truth assignment is acceptable if and only if for every clause, c_j, exactly one of l_{j1}, l_{j2}, l_{j3} is *true*, i.e., if exactly one of $m_{l_{j1}}, m_{l_{j2}}, m_{l_{j3}}$ is in the model, i.e., exactly one equation from E_j is in the model. Furthermore, every truth assignment assigns *true* to exactly one of u_i and $\bar{u}_i$, i.e., the corresponding model includes either m_{u_i} or $m_{\bar{u}_i}$, i.e., the corresponding model includes exactly one equation from F_i. Since a model is a causal model if and only if it includes exactly one equation from each E_j, exactly one equation from each F_i, and all the equations in G, it follows that a truth assignment is acceptable if and only if the corresponding model is a causal model. □

The second special case of the CAUSAL MODEL problem consists of those instances of the problem that satisfy the following two conditions: (a) the instance still has no propositional coherence constraints; and (b) each assumption class has exactly one model fragment. Since each assumption class contains exactly one model fragment, any causal model can be viewed as merely selecting a set of assumption classes. Hence, this special case identifies the second source of intractability: deciding which assumption classes to select is intractable. More abstractly, deciding which phenomena we want to model is itself intractable.

Lemma 5 *The* CAUSAL MODEL *problem is NP-hard even if its instances are required to satisfy the following conditions: (a) Equation 15; (b) $C = \emptyset$; and (c) every assumption class in $\mathcal{A}$ contains exactly one model fragment, i.e., the contradictory relation is empty.*

Proof: The proof of this lemma is a minor variation of the proof of Lemma 4. We use exactly the same reduction used there, except that we do not make m_l and $m_{\bar{l}}$ mutually *contradictory*, so that *contradictory* is the empty relation (as required). Instead, we introduce a set $Q = \{q_1, \ldots, q_n\}$ of additional parameters, a set $H = \{h_1, \ldots, h_n\}$ of additional equations, with

$$P(h_i) = P_c(h_i) = \{q_i\}, 1 \le i \le n$$

We then add equation h_i to both m_{u_i} and $m_{\bar{u}_i}$, so that any model that contains both m_{u_i} and $m_{\bar{u}_i}$ will be overconstrained. This is equivalent to making m_{u_i} and $m_{\bar{u}_i}$ mutually contradictory, so that the rest of the proof is identical. □

[11] This is true even if P_c were made identical to P (see [28]). However, the current definition of P_c makes this fact transparent.

An alternative view of the above two results is that a fundamental source of intractability is that a causal model is forced to choose between mutually contradictory model fragments. This was enforced in the first proof using the *contradictory* relation, and in the second using overconstrained sets of equations. This suggests that other ways of enforcing such a choice would also lead to intractability. In particular, if $\mathcal{C}$ contained constraints of the form $\neg m_1 \vee \neg m_2$ then any coherent model would have to choose at most one of m_1 or m_2, leading to intractability. The next case shows that other simple types of propositional coherence constraints also lead to intractability.

The third special case of the CAUSAL MODEL problem consists of those instances of the problem that satisfy the following conditions: (a) as in the first case, every causal model of the instance includes a model fragment from each assumption class; (b) model fragments in the same assumption class have the same sets of equations; and (c) $\mathcal{C}$ contains only definite horn clauses (a definite horn clause is a disjunction of literals with exactly one positive literal). Conditions (a) and (b) ensure that, if $\mathcal{C}$ were empty, then finding a causal model would be trivial: a causal model exists if and only if selecting an arbitrary model fragment from each assumption class leads to a causal model. Hence, the intractability of this problem stems from the causal model having to satisfy the constraints in $\mathcal{C}$, even when the constraints are restricted to be definite horn clauses.

Lemma 6 *The* CAUSAL MODEL *problem is NP-hard even if its instances are required to satisfy the following conditions: (a) Equation 15; (b) every causal model of the instance includes a model fragment from each assumption class; (c) model fragments in the same assumption class have the same sets of equations; and (d) $\mathcal{C}$ contains only definite horn clauses.*

Proof: Once again we use a reduction from an arbitrary instance (U, C) of the ONE-IN-THREE 3SAT problem. As in the earlier proofs, we introduce a model fragment m_l for each literal l, and make m_l and $m_{\bar{l}}$ mutually *contradictory.* Introduce the set $\mathcal{P} = \{p_0, \ldots, p_n\}$ of $(n+1)$ parameters, and the set $E = \{e_0, \ldots, e_n\}$ of $(n+1)$ equations. Let $p = p_0$ and $q = p_n$, and let

$$\begin{array}{ll} P(e_0) = \{p_0\} & P_c(e_0) = \{p_0\} \\ P(e_i) = \{p_{i-1}, p_i\} & P_c(e_i) = \{p_i\}, 1 \leq i \leq n \end{array}$$

Assign the equations to the model fragments as follows:

$$\begin{array}{c} E(m_{u_1}) = E(m_{\bar{u}_1}) = \{e_0, e_1\} \\ E(m_{u_i}) = E(m_{\bar{u}_i}) = \{e_i\}, 2 \leq i \leq n \end{array}$$

It is easy to see that the above assignment satisfies constraints (b) and (c) above. Finally, introduce the set $\mathcal{C} = \bigcup_{1 \leq j \leq m} \mathcal{C}_j$ of $3m$ propositional coherence constraints, where $\mathcal{C}_j$ contains three constraints derived from clause c_j. Let c_j contain literals l_{j1}, l_{j2}, l_{j3}. $\mathcal{C}_j$ contains the following three constraints:

$$\begin{aligned} \mathcal{C}_j = \{ & (m_{l_{j1}^-} \wedge m_{l_{j2}^-}) \equiv m_{l_{j3}}, \\ & (m_{l_{j1}^-} \wedge m_{l_{j3}^-}) \equiv m_{l_{j2}}, \\ & (m_{l_{j2}^-} \wedge m_{l_{j3}^-}) \equiv m_{l_{j1}} \ \} \end{aligned}$$

It is easy to verify that these constraints are equivalent to a set of definite horn clauses. Furthermore, in conjunction with the *contradictory* relation, they ensure that every causal model contains exactly one model fragment from $m_{l_{j1}}, m_{l_{j2}}, m_{l_{j3}}$. In fact, in conjunction with conditions (b) and (c) above, any model that selects a model fragment from each assumption class, and selects exactly one model fragment from each $m_{l_{j1}}, m_{l_{j2}}, m_{l_{j3}}$ is a causal model. But this means that a model is a causal model if and only if the corresponding truth assignment is acceptable (the corresponding truth assignment assigns literal l to be *true* if and only if m_l is in the model). □

An immediate consequence of any of the above three lemmas, in conjunction with Lemma 3 is that the CAUSAL MODEL problem is NP-complete.

Theorem 1 *The* CAUSAL MODEL *problem is NP-complete.*

Since a causal model exists if and only if a minimal causal model exists, the intractability of the CAUSAL MODEL problem immediately implies the intractability of the MINIMAL CAUSAL MODEL problem.

Theorem 2 *The* MINIMAL CAUSAL MODEL *problem is NP-hard.*

5 Causal approximations

The intractability of the MINIMAL CAUSAL MODEL problem implies that any algorithm for finding adequate models will be forced to search a significant portion of the exponentially large space of possible device models. Unfortunately, even for fairly simple devices, the space of of possible models is prohibitively large, making such a search unthinkable. However, this intractability of finding adequate models seems to directly contradict the informal observation that trained engineers are remarkably good at providing parsimonious causal explanations for phenomena. One way to resolve this apparent contradiction is to assume that trained engineers are not solving the general MINIMAL CAUSAL MODEL problem. Rather, the problem instances that they normally encounter are drawn from a subclass of the MINIMAL CAUSAL MODEL problem which can, in fact, be solved efficiently. In this section, we identify such an efficiently solvable subclass. We believe that commonly encountered instances of the MINIMAL CAUSAL MODEL problem are drawn from this subclass.

5.1 Upward failure property

Intuitively, the reason that the MINIMAL CAUSAL MODEL problem is intractable is that knowing whether a particular model is, or is not, a causal model tells us very little about which other models are, or are not, causal models. This means that there is no "clever" way to organize the search for adequate models, allowing us to rule out "large" parts of the search space by explicitly checking only a "small" part of the search space. With this intuition in mind, we introduce the upward failure property.

The upward failure property is based on the intuition that if a model is unable to explain the phenomenon of interest, there is little reason to believe that a simpler model will be able to explain that phenomenon. We make this precise with the following definition, which is similar in spirit to the one given in [44]:

Definition 11 (Upward failure property) *An instance $\mathcal{I}$ of the* MINIMAL CAUSAL MODEL *problem is said to satisfy the upward failure property if and only if for all coherent models $M \subseteq \mathcal{M}$, if M is not a causal model, then no strictly simpler model is a causal model, i.e., no model $M' \subseteq \mathcal{M}$ and $M' < M$ is a causal model.*

In essence, the upward failure property property says that the simpler the model, the less it can explain. Of course, it is by no means obvious that simpler models explain fewer phenomena. However, it does seem to be standard engineering practice that models that account for more phenomena are more complex by our definition, i.e., modeling more phenomena more accurately leads to models that can explain more. This is, of course, not an argument for claiming that the upward failure property is satisfied by all commonly encountered instances of the MINIMAL CAUSAL MODEL problem. Rather, it merely provides a motivation for our definition of the upward failure property.

5.1.1 Finding a minimal causal model

If an instance, $\mathcal{I}$, of the MINIMAL CAUSAL MODEL problem satisfies the upward failure property, a causal model can be simplified to a minimal causal model using the function *find-minimal-causal-model* shown in Figure 10. This function takes two arguments: (a) $\mathcal{I}$; and (b) a coherent model M. It returns a minimal causal model that is simpler than M. If there is more than one such minimal causal model, it returns the first one it finds. If no such model exists, it returns nil.

The *simplifications* function, used in *find-minimal-causal-model*, when applied to a coherent model M, returns the set of coherent models that are immediate simplifications of M. A coherent model M' is an immediate simplification of M if and only if $M' < M$ and there does not exist a coherent model M'' such that $M' < M'' < M$.

$$\begin{aligned} &\mathit{simplifications}(M, \mathcal{I}) = \\ &\quad \{M' :\ M' \text{ is coherent wrt } \mathcal{I} \wedge M' < M \\ &\qquad\quad \wedge\ (\forall M'')\ M' < M'' < M \Rightarrow M'' \text{ is not coherent wrt } \mathcal{I}\} \end{aligned} \tag{16}$$

```
function find-minimal-causal-model($\mathcal{I}, M$)
  /* $\mathcal{I}$ is assumed to satisfy the upward failure property */
  /* $M$ is assumed to be coherent */
  if $M$ is not a causal model then
    /* Since no simpler model can be a causal model */
    return nil
  else
    for each $M' \in$ simplifications($M, \mathcal{I}$) do
      result := find-minimal-causal-model($\mathcal{I}, M'$)
      if result $\neq$ nil then
        /* A simpler causal model has been found */
        return result
      endif
    endfor
    /* No simplification is a causal model, but $M$ is */
    return $M$
  endif
end
```

Figure 10: Function *find-minimal-causal-model*

Find-minimal-causal-model($\mathcal{I}, M$) works by systematically searching the immediate simplifications of M, until it finds a causal model M' such that all the immediate simplifications of M' are not causal models. The upward failure property then assures us that M' is a minimal causal model. The following lemma establishes the correctness of this function:

Lemma 7 *Let $\mathcal{I}$ be an instance of the* MINIMAL CAUSAL MODEL *problem that satisfies the upward failure property, and let $M \subseteq \mathcal{M}$ be a coherent model. Then find-minimal-causal-model($\mathcal{I}, M$) returns an adequate model (i.e., a minimal causal model) of $\mathcal{I}$ that is simpler than M, if it exists, and nil otherwise.*

Proof: We prove this lemma by induction. There are two base cases: (a) if M is not a causal model, then the upward failure property guarantees that there are no causal models simpler than M, and the function correctly returns nil; and (b) if M is a causal model with no immediate simplifications, then the function correctly returns M. The inductive step also has two cases: (a) if the recursive call to every immediate simplification of M returns nil, then the inductive hypothesis tells us that there is no causal model strictly simpler than M, and since M is a causal model, the function correctly returns M; and (b) if the recursive call to some immediate simplification, M', of M returns a model, the inductive hypothesis ensures that this model is a minimal causal model that is simpler than M', and hence simpler than M, and hence the function correctly returns this model. □

The following lemma states that if the immediate simplifications of a coherent model can be computed in polynomial time, then *find-minimal-causal-model* also runs in polynomial time:

Lemma 8 *Let $\mathcal{I}$ be an instance of the* MINIMAL CAUSAL MODEL *problem, and let $M \subseteq \mathcal{M}$ be a coherent model. If the immediate simplifications of every coherent model of $\mathcal{I}$ can be computed in time polynomial in the size of $\mathcal{I}$, then find-minimal-causal-model($\mathcal{I}, M$) terminates in time polynomial in the size of $\mathcal{I}$.*

Proof: Excluding the recursive calls, the only significant work done by *find-minimal-causal-model($\mathcal{I}, M$)* is to check if M is a causal model, and to generate the immediate simplifications of M (if necessary). Lemma 3 tells us that the former can be done in polynomial time, while the latter can be done in polynomial time by assumption. Hence, we need only show that there are a polynomial number of recursive calls to *find-minimal-causal-model.* One can see that if *find-minimal-causal-model* returns nil, it makes no recursive calls. Hence, one can verify that, of the recursive calls made by *find-minimal-causal-model,* at most one can itself make recursive calls. Hence, in conjunction with the fact that every call to *find-minimal-causal-model* can make at most a polynomial number of recursive calls, it follows that the total number of recursive calls is polynomial if and only if the maximum depth of the recursion is polynomial. Now, every recursive call made by *find-minimal-causal-model($\mathcal{I}, M$)* replaces the model M by an immediate simplification of M, constructed by dropping and/or approximating some set of model fragments in M. Hence, once a model fragment, m, is removed from M, no deeper recursive call uses a model containing m. Hence, the depth of the recursion is bounded by the number of model fragments in $\mathcal{M}$. Hence, the total number of recursive calls is polynomial. □

To find a minimal causal model of $\mathcal{I}$, we merely invoke the function *find-minimal-causal-model* on each of the most accurate coherent models of $\mathcal{I}$. The most accurate coherent models of $\mathcal{I}$ are those coherent models that are not strictly simpler than any other coherent models:

$$\{M : M \text{ is coherent } \wedge \neg\exists M' \; M' \text{ is coherent } \wedge M < M'\} \tag{17}$$

Since every minimal causal model of $\mathcal{I}$ must be simpler than (though not necessarily strictly simpler than) some most accurate coherent model of $\mathcal{I}$, it is easy to see that a minimal causal model can be identified by systematically invoking *find-minimal-causal-model* on each of the most accurate models of $\mathcal{I}$. Hence, we have the following:

Theorem 3 *Let $\mathcal{I}$ be an instance of the* MINIMAL CAUSAL MODEL *problem that satisfies the upward failure property. If the most accurate coherent models of $\mathcal{I}$ can be* *generated in time polynomial in the size of $\mathcal{I}$, and if the immediate simplifications of any coherent model of $\mathcal{I}$ can be generated in time polynomial in the size of $\mathcal{I}$. then a minimal causal model of $\mathcal{I}$ can be found in time polynomial in the size of $\mathcal{I}$.*

Proof: Immediate consequence of Lemmas 7 and 8 and the above discussion. □

5.1.2 Discussion

We have seen that the upward failure property is useful because it leads to an efficient algorithm for finding an adequate model. However, it has a major drawback: it is very difficult to decide whether or not a particular instance of the MINIMAL CAUSAL MODEL problem satisfies the upward failure property. For example, a straightforward use of Definition 11 requires us to check every model in the space of possible models. Since the space of possible models is exponentially large, any such check is unthinkable. In fact, the upward failure property was suggested as a way around having to check the whole space of possible models. Unfortunately, it does not seem to have succeeded in helping us to circumvent this problem.

This drawback of the upward failure property stems from the fact that it is a *global* property, i.e., a property of the whole space of possible models. What we want is a *local* property that entails the upward failure property, i.e., a property of the encoding of $\mathcal{I}$ that can be checked efficiently, that will ensure that $\mathcal{I}$ satisfies the upward failure property.

We now present some local properties of $\mathcal{I}$ that ensure that the conditions of Theorem 3 are satisfied. In particular, we will go back to the sources of intractability identified in the previous section, and place appropriate restrictions on $\mathcal{I}$: (a) we will introduce a new class of *approximations*, called *causal approximations*, that will address the problem of selecting model fragments from selected assumption classes; (b) we will add additional constraints to $\mathcal{C}$, called *ownership constraints*, that will address the problem of selecting assumption classes; and (c) we will restrict the expressive power of constraints in $\mathcal{C}$.

5.2 Preliminary restrictions

We start by introducing three preliminary restrictions on $\mathcal{I}$. First, we assume that the *contradictory* relation partitions the set of model fragments into the set of assumption classes (Equation 15). This assumption is based on the intuition that there is little reason for descriptions of different phenomena to be mutually contradictory.

Second, we assume that each assumption class has a single, most accurate model fragment:

$$(\forall A \in \mathcal{A})(\exists m \in A)(\forall m' \in A)\ m \neq m' \Rightarrow \mathit{approximation}(m, m') \qquad (18)$$

In other words, we assume that each phenomena has a single best description. This is a reasonable assumption as long as we only model fairly well understood phenomena, i.e., where there is broad consensus amongst the domain experts about how best to model the phenomena.

Note that the above restriction appears to be a problem when a given phenomena can be modeled with multiple ontologies. In such cases, it may not be possible to say that one ontology is more accurate than the other, leading to multiple most accurate model fragments in an assumption class. However, this does not pose a problem if the different ontologies are not mutually contradictory, so that model fragments that use different ontologies are in different assumption classes. This is often the case, since different ontologies are often used for different purposes.

An important consequence of the above restriction is that it leads to $\mathcal{I}$ having a single most accurate model: the most accurate model of $\mathcal{I}$ is just the set of most accurate model fragments of the assumption classes of $\mathcal{I}$. This brings us to our third assumption: we assume that the most accurate model of $\mathcal{I}$ is coherent. This ensures that $\mathcal{I}$ has exactly one most accurate coherent model, and this model can be generated in polynomial time.

5.3 Causal Approximations

The first source of intractability, i.e., having to choose model fragments from selected assumption classes, is addressed by the introduction of *causal approximations*. The basic idea underlying the definition of causal approximations is that more approximate descriptions often tend to involve fewer parameters. Furthermore, more approximate descriptions tend to explain less about a phenomenon than more accurate descriptions.

For example, Figure 5 showed different descriptions of electrical conduction, and Figure 6 showed the *approximation* relation between these descriptions. Note that the parameters in the equations of the more approximate descriptions ($V_w = 0$ and $i_w = 0$) are a subset of the parameters in the equations of the more accurate description ($V_w = i_w R_w$). Furthermore, only `Resistor(wire-1)` is able to explain the relationship between V_w, i_w, and R_w. In the following, we make the above idea precise, and investigate its consequences.

5.3.1 Definitions

A *local* parameter is a parameter that can be causally determined only by equations of model fragments in a single assumption class.

Definition 12 (Local parameters) *A parameter p is said to be local to a model fragment $m \in \mathcal{M}$ if and only if p can be causally determined by the equations of m, but not by the equations of any model fragment that does not contradict m:*

$$p \in P_c(m) \wedge (\forall m' \in \mathcal{M})\ m \neq m' \wedge p \in P_c(m') \Rightarrow \mathit{contradictory}(m, m')$$

A parameter is said to be shared if it is not local to any model fragment.

The above is used in the following definition of causal approximations. The idea underlying this definition is that if m_2 is a causal approximation of m_1, then any causal orientation of the equations of m_2 can be extended to a causal orientation of the equations of m_1, such that the latter causal orientation entails a superset of causal relations, i.e., m_1 can explain more than m_2:

Definition 13 (Causal approximations) *A model fragment m_2 is said to be a causal approximation of a model fragment m_1 if and only if:*

1. *m_2 is an approximation of m_1;*

2. *There exists a 1-1 mapping $G : m_2 \rightarrow m_1$ such that for each $e \in m_2$, $P(e) \subseteq P(G(e))$, and $P_c(e) \subseteq P_c(G(e))$. G is called a correspondence mapping, and e and $G(e)$ are said to be corresponding equations; and*

3. *Let E^* denote the equations of m_1 that have no corresponding equations in m_2, and let P^* denote the set of parameters that are local to m_1, but not local to m_2. Then there exists an onto causal mapping $L : E^* \rightarrow P^*$. L is called a local causal mapping with respect to correspondence mapping G.*

Condition 1 ensures that causal approximations are approximations. Condition 2 ensures that for any causal orientation of an equation $e \in m_2$, there exists a causal orientation of $G(e) \in m_1$ which entails a superset of causal relations. Condition 3 ensures that additional equations in m_1 can be oriented to causally determine newly introduced local parameters.

For example, the approximation relation between `Temperature-dependent-resistance(wire-1)` and `Constant-resistance(wire-1)` shown in Figure 11 is a causal approximation if we assume that R_{w0}, α_w, and T_{w0} are local parameters of `Temperature-dependent-resistance(wire-1)`.

$$
\begin{array}{rl}
\texttt{Constant-resistance(wire-1):} & \{exogenous(\boldsymbol{R}_w)\} \\
\texttt{Temperature-dependent-resistance(wire-1):} & \{\boldsymbol{R}_w = R_{w0}(1 + \alpha_w(T_w - T_{w0})), \\
 & \;exogenous(\boldsymbol{R}_{wo}), \\
 & \;exogenous(\boldsymbol{\alpha}_w), \\
 & \;exogenous(\boldsymbol{T}_{wo})\}
\end{array}
$$

$$
\begin{array}{c}
approximation(\texttt{Temperature-dependent-resistance(wire-1)}, \\
\texttt{Constant-resistance(wire-1)})
\end{array}
$$

Figure 11: Model fragments describing a wire's resistance.

In particular, $exogenous(\boldsymbol{R}_w)$ and $\boldsymbol{R}_w = R_{w0}(1 + \alpha_w(T_w - T_{w0}))$ are corresponding equations, and the local causal mapping, L, with respect to this correspondence mapping is:

$$L(exogenous(\boldsymbol{R}_{wo})) \;=\; R_{w0}$$

$$\begin{aligned} L(exogenous(\boldsymbol{\alpha}_w)) &= \alpha_w \\ L(exogenous(\boldsymbol{T}_{wo})) &= T_{w0} \end{aligned}$$

One can show that the causal approximation relation between model fragments is transitive. Hence, to check that all approximations are causal approximations, it is sufficient to check that the immediate approximations of each model fragment are causal approximations.

It is worth noting that the restriction that local parameters in a model fragment cannot be causally determined by equations of model fragments in other assumption classes is not a serious one. It is easy to convert a local parameter into a shared parameter by defining a new assumption class. For example, to convert α_w (Figure 11) into a shared parameter, we would (a) define a new assumption class with one model fragment $m = \{exogenous(\boldsymbol{\alpha}_w)\}$; and (b) remove $exogenous(\boldsymbol{\alpha}_w)$ from the equations of `Temperature-dependent-resistance(wire-1)`. After this conversion, α_w is not necessarily local to any assumption class.

5.3.2 Causal approximations and the upward failure property

Causal approximations plays a key role in ensuring that the upward failure property is satisfied. The following theorem tells us that when all approximations are causal approximations, the causal relations entailed by a model decrease monotonically as we simplify models without dropping assumption classes. This means that if a model does not explain the expected behavior, then a simpler model that uses the same assumption classes also does not explain the expected behavior. It is easy to see that this is just a restricted version of the upward failure property.

Theorem 4 *Let $\mathcal{I}$ be an instance of* MINIMAL CAUSAL MODEL *such that all approximations are causal approximations, and the contradictory relation partitions the set $\mathcal{M}$ of model fragments into the set $\mathcal{A}$ of assumption classes. Let $M_1, M_2 \subseteq \mathcal{M}$ be coherent models such that M_1 and M_2 contain model fragments from the same assumption classes, and $M_2 \leq M_1$. The causal relations entailed by the equations of M_2 are a subset of the causal relations entailed by the equations of M_1, i.e., $C(E(M_2)) \subseteq C(E(M_1))$.*

Proof: Let $F_2 : E(M_2) \rightarrow P(M_2)$ be any onto causal mapping. We construct an onto causal mapping $F_1 : E(M_1) \rightarrow P(M_1)$ such that $C_{F_2} \subseteq C_{F_1}$. For the equations of each model fragment $m \in M_1$, define F_1 as follows.

1. If $m \in M_2$, then for each equation in m, define F_1 to be the same as F_2.

2. Otherwise, there exists a unique $m' \in M_2$ such that m' is a causal approximation of m. Let G be the correspondence mapping between the equations of m' and m and let L be the local causal mapping. For each equation $e \in m$, if there is an equation $e' \in m'$ such that $G(e') = e$, then let $F_1(e) = F_2(e')$. Otherwise, let

$F_1(e) = L(e)$. F_1 is well defined on each equation in m_1 by Conditions 2 and 3 in Definition 13. Using Condition 2 in Definition 13, one can also verify that the direct causal dependencies entailed by F_1 restricted to m are a superset of the direct causal dependencies entailed by F_2 restricted to m'.

It is easy to see that F_1 is well defined, since if $m_1, m_2 \in M_1$ are any two distinct model fragments, the range of F_1 restricted to equations in m_1 is disjoint from the range of F_1 restricted to equations in m_2. Since F_1 is defined for all equations in $E(M_1)$, and M_1 is coherent, it follows that F_1 is an onto causal mapping. Finally, using the last sentence in point 2 above, it is easy to see that $C_{F_2} \subseteq C_{F_1}$. □

The above theorem can be illustrated graphically by considering the model resulting from replacing `Constant-resistance(wire-1)` by `Temperature-dependent-resistance(wire-1)` in the model shown in Figure 7. The direct causal dependencies generated by a causal mapping on the equations of this model are shown in Figure 12. In comparing this figure with the direct causal dependencies in Figure 9 one can see that they are identical except (a) R_w is now determined by the equation $\boldsymbol{R_w} = R_{w0}(1 + \alpha_w(T_w - T_{w0}))$ instead of the equation $exogenous(\boldsymbol{R_w})$; and (b) the additional parameters, R_{w0}, α_w, and T_{w0}, are determined by the additional equations, $exogenous(\boldsymbol{R_{w0}})$, $exogenous(\boldsymbol{\alpha_w})$, and $exogenous(\boldsymbol{T_{w0}})$, respectively. One can easily verify that the causal dependencies have increased monotonically with this replacement.

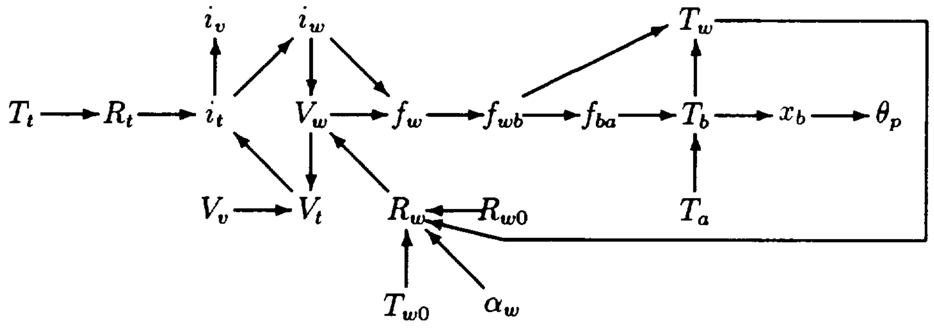

Figure 12: The direct causal dependencies resulting from replacing `Constant-resistance(wire-1)` by `Temperature-dependent-resistance(wire-1)` in the model of Figure 7

As a consequence of the above theorem, if a coherent model does not explain the expected behavior, it follows that no simpler coherent model that uses the same set of assumption classes can explain the expected behavior. Hence, when all approximations are causal approximations, a restricted version of the upward failure property is satisfied. Note that, unlike the upward failure property, it is easy to decide whether or not all approximations are causal approximations.

Causal approximations are particularly useful because they are commonly found in modeling the physical world. Table 1 shows a list of commonly used approximations, all of which are causal approximations.

Inertialess objects	Rigid bodies
Inviscid flow	Elastic collisions
Frictionless motion	Ideal gas law
Zero or constant gravity	Ideal heat engines
Non-relativistic mass and motion	No thermal expansion
Ideal thermal insulators and conductors	Constant thermal conductance
Ideal electrical insulators and conductors	Constant resistance and resistivity

Table 1: Examples of causal approximations

The details of the above causal approximations, including the actual equations used, can be found in Appendix A. The ubiquity of causal approximations suggests that we have identified an important property of commonly occurring instances of the MINIMAL CAUSAL MODEL problem.

5.4 Selecting assumption classes

While the use of causal approximations addresses the first source of intractability, it does not address the second source (i.e., the problem of selecting assumption classes). A simple example illustrates that causal approximations alone are not sufficient to ensure that the causal relations decrease monotonically as models are simplified by dropping model fragments. Let $A_1 = \{m_{11}, m_{12}\}$ and $A_2 = \{m_2\}$ be assumption classes, and let the equations of model fragments m_{11}, m_{12}, and m_2 be defined as follows:

$$m_{11} = \{\boldsymbol{x} = \boldsymbol{y}, \boldsymbol{y} = \boldsymbol{z}\};\; m_{12} = \{\boldsymbol{x} = \boldsymbol{y}, \mathit{exogenous}(\boldsymbol{y})\};\; m_2 = \{\mathit{exogenous}(\boldsymbol{x})\}$$

Furthermore, let m_{12} be an approximation of m_{11}. It is easy to verify that m_{12} is a causal approximation of m_{11}. Let $M_1 = \{m_{11}, m_2\}$ and $M_2 = \{m_{12}\}$ be two models. Assuming that there are no propositional coherence constraints, it is easy to verify that both M_1 and M_2 are coherent models, and that $M_2 < M_1$. However, y causally depends on x in the causal ordering generated from M_1, while x causally depends on y in the causal ordering generated from M_2. Hence, in simplifying M_1 to M_2, the causal relations have not decreased monotonically, and the upward failure property is not satisfied.

Intuitively, the problem appears to be that, M_2 does not include all phenomena that were possibly "relevant" to its parameters. In particular, M_2 used the parameter x, but did not include m_2, even though an equation in m_2 could causally determine x. We use this intuition to ensure that the causal relations decrease monotonically even when models are simplified by dropping assumption classes. We formalize this intuition by defining a set of *ownership constraints* that will ensure that coherent models include all possibly "relevant" phenomena.

The parameters *owned* by an assumption class are the parameters that can be causally determined by some equation of some model fragment in the assumption class.

Definition 14 (Parameter ownership) *The parameters owned by an assumption class A, denoted by $owns(A)$, are the parameters that can be causally determined by the equations of model fragments of A:*

$$owns(A) = \bigcup_{m \in A} P_c(m)$$

One can view an assumption class as being possibly "relevant" to the parameters that it *owns*. We ensure that coherent models will contain model fragments from all possibly "relevant" assumption classes, by adding constraints of the form

$$m \Rightarrow A$$

to the set $\mathcal{C}$ of propositional coherence constraints, whenever assumption class A *owns* a parameter that can be causally determined by an equation in m, i.e., when $P_c(m) \cap owns(A)$ is not empty. This will ensure that whenever a coherent model contains model fragment m, it will also contain a model fragment from A. We call the above set of constraints *ownership constraints*:

Definition 15 (Ownership constraints) *Let $\mathcal{I}$ be an instance of the* MINIMAL CAUSAL MODEL *problem. The set $\mathcal{O}$ of ownership constraints of $\mathcal{I}$ are defined as follows:*

$$\mathcal{O} = \{m \Rightarrow A : m \in \mathcal{M} \ \wedge \ A \in \mathcal{A} \ \wedge P_c(m) \cap owns(A) \neq \emptyset\}$$

When $\mathcal{C}$ contains all the ownership constraints we can extend Theorem 4 to all coherent models.

Theorem 5 *Let $\mathcal{I}$ be an instance of* MINIMAL CAUSAL MODEL *such that all approximations are causal approximations, and the contradictory relation partitions the set $\mathcal{M}$ of model fragments into the set $\mathcal{A}$ of assumption classes. Let $\mathcal{C}$ contain all the ownership constraints of $\mathcal{I}$. Let $M_1, M_2 \subseteq \mathcal{M}$ be coherent models such that $M_2 \leq M_1$. The causal relations entailed by the equations of M_2 are a subset of the causal relations entailed by the equations of M_1, i.e., $C(E(M_2)) \subseteq C(E(M_1))$.*

Proof: Let $F_1 : E(M_1) \rightarrow P(M_1)$ and $F_2 : E(M_2) \rightarrow P(M_2)$ be any onto causal mappings. Using F_1 and F_2 we construct an onto causal mapping $F : E(M_1) \rightarrow P(M_1)$ such that $C_{F_2} \subseteq C_F$. Let us partition M_1 into two mutually disjoint sets M_{11} and M_{12} such that M_{11} and M_2 have no model fragments from the same assumption classes, while M_{12} and M_2 have model fragments from the same assumption classes.

Since C contains all the ownership constraints, it follows that the range of F_1 restricted to $E(M_{11})$ is disjoint from the range of F_2 (otherwise an equation in $E(M_2)$ could causally determine a parameter owned by an assumption class not used in M_2).

Define F as follows. For each equation in $E(M_{11})$ let F be identical to F_1. Define F on the equations in $E(M_{12})$ in exactly the same way as was done in the proof of Theorem 4, i.e., causally orient corresponding equations in the same way, and orient the remaining equations according to the local causal mappings. F is well defined because the last sentence in the above paragraph ensures that the ranges of F restricted to $E(M_{11})$ and F restricted to M_{12} are disjoint. F is an onto causal mapping because F is defined on every equation in $E(M_1)$ and M_1 is complete. As in the proof of Theorem 4, it is easy to see that the direct causal dependencies entailed by F_2 are a subset of the direct causal dependencies entailed by F restricted to M_{12}, and hence $C_{F_2} \subseteq C_F$. □

To illustrate the above theorem, consider the model fragment shown in Figure 13, which describes the electromagnetic field generated in the wire. ($M{+}(\mu_w, i_w)$ is a qualitative equation stating that the magnetic moment, μ_w, is proportional to the current, i_w.) Now consider the model resulting from adding `Electromagnet(wire-1)` to the model shown in Figure 7. Since none of the model fragments in the original model can causally determine μ_w, there are no relevant ownership constraints, and the above theorem tells us that the resulting model entails a superset of causal dependencies. Figure 14 shows the direct causal dependencies entailed by the resulting model, and a comparison with with Figure 9 confirms the theorem.

`Electromagnet(wire-1)`:$\{M{+}(\mu_w, i_w)\}$

Figure 13: Model fragment describing electromagnetism

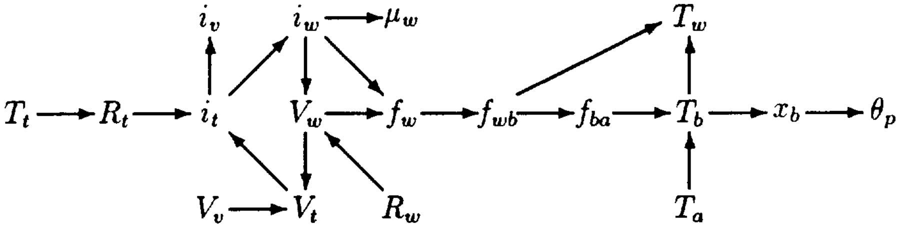

Figure 14: The direct causal dependencies resulting from adding `Electromagnet(wire-1)` to the model in Figure 7.

How reasonable are the ownership constraints? While they appear quite restrictive, under certain circumstances we get them for free. In particular, consider the situation in which each equation can causally determine exactly one parameter. This situation is found in QP Theory [15] and its derivatives, e.g., [14]. In this case all the parameters are local to some assumption class, and hence no model fragment can causally determine a parameter owned by a different assumption class. Hence, there are *no* ownership constraints!

Unfortunately, as we have argued earlier, the constraint that each equation can causally determine exactly one parameter is also restrictive. In the absence of this constraint, the ownership constraints appear to be necessary to guarantee that the upward failure property is satisfied.

5.5 Generating all immediate simplifications

Having established local conditions under which the upward failure property is satisfied, we now turn to the other element of our efficient model selection algorithm: efficiently generating the immediate simplifications of a coherent model. The complexity of generating the immediate simplifications of a coherent model is critically dependent upon the expressive power of the constraints in $\mathcal{C}$. In fact, a consequence of Lemma 6 is that if $\mathcal{C}$ contains definite horn clauses, then the immediate simplifications of coherent models cannot be generated efficiently.

In [28] we show that a minimal causal model can be found efficiently if we restrict the constraints in $\mathcal{C}$ to have the following form:

$$m_1 \wedge m_2 \wedge \ldots \wedge m_n \Rightarrow A \tag{19}$$

where $m_1, m_2, \ldots, m_n$ are model fragments, and A is an assumption class. However, this restriction does not ensure that the immediate simplifications of a coherent model can be generated efficiently. Hence, the efficient algorithm discussed in [28] is a modification of the function *find-minimal-causal-model.* In this paper, in the interests of brevity, we will not discuss this modification. Rather, we will restrict the constraints in $\mathcal{C}$ to have the following form:

$$m \Rightarrow A \tag{20}$$

where m is a model fragment and A is an assumption class. (Note that the ownership constraints have this form.) We now show that this restriction will ensure that the immediate simplifications of a coherent model can be generated efficiently.

Informally, a model can be simplified either by approximating one of its model fragments, or by dropping one of its model fragments (or by some combination of approximating and dropping model fragments). Consider, first, simplifying a model by approximating one of its model fragments. An *acceptable approximation* of a model fragment $m \in M$ is an *approximation*, m', of m such that the model resulting from replacing m by m' in M satisfies all the constraints in $\mathcal{C}$. (However, notice that the resulting model need not be coherent.)

Definition 16 (Acceptable approximation) *Let M be a coherent model, and let $m \in M$ be any model fragment. A model fragment m' is an acceptable approximation of m with respect to M if and only if (a) approximation(m, m'); and (b)*

$(M \setminus \{m\}) \cup \{m'\}$ satisfies all the constraints in $\mathcal{C}$. An immediate acceptable approximation of m with respect to M is an acceptable approximation, m', such that there is no acceptable approximation m'' with approximation(m'', m').

Given that the constraints in $\mathcal{C}$ are restricted as in Equation 20, it is easy to see that if m' is an *approximation* of m, but m' is not an acceptable approximation of m with respect to M, then no coherent model simpler than M can contain m'. A *type 1 simplification* of a coherent model, M, is a coherent model $M' < M$, where M' is the result of replacing a model fragment $m \in M$ by one of its immediate acceptable approximations.

Definition 17 (Type 1 simplification) *Let M be a coherent model, and let $m \in M$ be any model fragment. Let m' be an immediate acceptable approximation of m with respect to M. $(M \setminus \{m\}) \cup \{m'\}$ is a type 1 simplification of M if and only if it is coherent.*

It is easy to verify that a type 1 simplification of M is an immediate simplification of M, i.e., is an element of *simplifications*$(M, \mathcal{I})$. Furthermore, we have:

Lemma 9 *The type 1 simplifications of M can be generated in polynomial time.*

Proof: Immediate from the definitions and Lemma 3. □

Next, we consider simplifying a model by dropping model fragments. The following lemma is the basis for identifying conditions under which model fragments can be dropped. It tells us that if the result of replacing a model fragment, $m \in M$, by one of its immediate acceptable approximations, m', does not lead to a type 1 simplification of M, i.e., the resulting model is not coherent, then no model simpler than M can contain m' or any of its approximations.

Lemma 10 *Let M be a coherent model, and let $m \in M$ be any model fragment. Let m' be an immediate acceptable approximation of m with respect to M, and let $M' = (M \setminus \{m\}) \cup \{m'\}$. If M' is not coherent, then every model strictly simpler than M that contains either m' or any of its approximations, is also not coherent.*

Proof: Let $M_2 < M$ be a model such that M_2 contains a model fragment m_2 that is either m' or is an *approximation* of m'. Assuming that M_2 is coherent, we show that M' is also coherent. Let $F_2 : E(M_2) \rightarrow P(M_2)$ be any onto causal mapping. Let $F : E(M_1) \rightarrow P(M_1)$ be an onto causal mapping constructed from F_2 in the same way that F was constructed from F_2 in the proof of Theorem 5. Construct an onto causal mapping $F' : E(M') \rightarrow P(M')$ as follows. Let F' be identical to F on the equations of model fragments common to M' and M. That leaves only the equations in m'. Define F' on the equations in m' using F_2 on the equations in m_2, by orienting corresponding equations in the same way and using the local causal mapping for the remaining equations (similar to the method used in the proof of Theorem 4). One can see that F' is well defined for every equation in $E(M')$. Hence, $|E(M')| \leq |P(M')|$. However, the only difference between M' and M is that m' is replaced by m. Since m' is a causal approximation of m, this replacement adds at least as many parameters as equations, and hence $|E(M)| \leq |P(M)|$. However, M is coherent, so that $|E(M)| = |P(M)|$. Hence, $|E(M')| = |P(M')|$ so that M' is coherent. □

Hence, if no type 1 simplification of M contains an immediate acceptable approximation of m, it follows that any immediate simplification of M not containing m was constructed by dropping m from M (and not by replacing m by an immediate acceptable approximation). Model fragments like m are called *removable* model fragments.

Definition 18 (Removable model fragment) *Let M be a coherent model, and let $m \in M$ be a model fragment. m is said to be removable from M if and only if replacing m by any of its immediate acceptable approximations results in a model that is not coherent.*

Using the above definition and the preceding comments, we define the type 2 simplifications, which are the immediate simplifications of a coherent model generated by dropping a set of model fragments.

Definition 19 (Type 2 simplifications) *Let M be a coherent model. Then $M' \subset M$ is a type 2 simplification of M if and only if (a) M' is coherent; (b) every model fragment in $M \backslash M'$ is removable from M; and (c) there is no coherent model $M'' \subset M$ such that $M' \subset M''$.*

Using Lemma 10, it is easy to see that the type 2 simplifications are also immediate simplifications. We will shortly show how the type 2 simplifications of a coherent model can be generated efficiently. Before doing that, we show that every immediate simplification of a coherent model is either a type 1 simplification or a type 2 simplification.

Lemma 11 *Let M be a coherent model, and let $M' \in$ simplifications$(M, \mathcal{I})$. Then M' is either a type 1 or a type 2 simplification of M.*

Proof: Assume that M' is not a type 1 or a type 2 simplification. Clearly $M' \not\subset M$, for otherwise a model fragment in $M \backslash M'$ is not removable from M, in which case M' is not an immediate simplification. (If all model fragments in $M \setminus M'$ are removable, then M' must be simpler than a type 2 simplification.) Hence, there are model fragments $m' \in M'$ and $m \in M$ such that *approximation*(m, m'). Let m'' be an immediate acceptable approximation of m with respect to M such that either m'' is the same as m', or *approximation*(m'', m'), and let $M'' = (M \setminus \{m\}) \cup \{m''\}$. Such an m'' must exist, and M'' must be coherent, for otherwise Lemma 10 tells us that M' is not coherent. Clearly, M'' is a type 1 simplification of M, and $M' \leq M''$. Since M' is not a type 1 simplification of M, it follows that M' is not an immediate simplification of M. □

To generate the type 2 simplifications of a coherent model we need to remove minimal sets of removable model fragments. To identify such minimal sets of model fragments, we introduce *remove-before*$_M$, a binary relation on model fragments defined with respect to a coherent model M. Intuitively, *remove-before*$_M(m_1, m_2)$ says that model fragment m_1 must be removed from M before (or with) model fragment m_2.

Definition 20 (Remove-before) *Let M be a coherent model, and let $m_1, m_2 \in M$ be model fragments of M. Let $F : E(M) \rightarrow P(M)$ be any onto causal mapping. remove-before$_M(m_1, m_2)$ is true, if and only if one of the following is satisfied: (a) the assumption class of m_2 is A_2 and $\mathcal{C}$ contains a constraint $m_1 \Rightarrow A_2$; or (b) m_1 contains an equation e_1 and m_2 contains an equation e_2, such that $F(e_2) = p_2$ and $p_2 \in P(e_1)$.*

Condition (a) includes dependencies between model fragments stemming from the constraint in $\mathcal{C}$: because of the constraint $m_1 \Rightarrow A_2$, m_2 cannot be removed from M before m_1. Condition (b) includes dependencies stemming from the causal ordering: removing m_2 before removing m_1 will result in an incomplete set of equations (intuitively, p_2 will not be causally determined). As in Definition 2, the identity of the causal mapping used does not matter, since we will be interested in the transitive closure of *remove-before*$_M$. The importance of the *remove-before*$_M$ relation is embodied in the following lemma:

Lemma 12 *Let M be a coherent model. $M' \subset M$ is a coherent model if and only if for every $m_1 \in M'$, if remove-before$_M(m_1, m_2)$ then $m_2 \in M'$.*

Proof: M' is a coherent model if and only if (a) all the constraints in $\mathcal{C}$ are satisfied; and (b) there is an onto causal mapping $F' : E(M') \rightarrow P(M')$. It is easy to verify that M' satisfies all the constraints in $\mathcal{C}$ if and only if all the *remove-before*$_M$ constraints stemming from $\mathcal{C}$ are satisfied. Since $M' \subset M$, F' can be obtained by restricting any onto causal mapping $F : E(M) \rightarrow P(M)$ to the equations in $E(M')$. By using the F used in defining *remove-before*$_M$ (Definition 20), it is easy to verify that F' is onto if and only if all the *remove-before*$_M$ constraints stemming from F are satisfied. □

The above lemma provides us with an algorithm for finding all type 2 simplifications of a coherent model M. This algorithm is shown in Figure 15.

The algorithm proceeds by constructing a directed graph from the *remove-before*$_M$ relation. The strongly connected components of this graph correspond to sets of model fragments that must be removed simultaneously from M, i.e., m_1 and m_2 are in the

```
function find-type-2-simplifications($M, \mathcal{I}$)
  /* Find all type 2 simplifications of a coherent model $M$ */
  type-2-simplifications ← nil
  Construct a directed graph, $G$, whose nodes are the model fragments in $M$
    and which has an edge from $m_1$ to $m_2$ iff remove-before$_M(m_1, m_2)$
  Find the strongly connected components of $G$
  for every strongly connected component, $C$, that has no edges entering it do
    /* $C$ is a minimal set of model fragments that can be removed from $M$ */
    /* to ensure that the resulting model is coherent. */
    if all the model fragments in $C$ are removable from $M$ then
      Add $(M \setminus C)$ to type-2-simplifications
    endif
  endfor
  return type-2-simplifications
end
```

Figure 15: Function *find-type-2-simplifications*

same strongly connected components if the transitive closure of the *remove-before*$_M$ relation implies both *remove-before*$_M(m_1, m_2)$ and *remove-before*$_M(m_2, m_1)$. Hence, a strongly connected component C that has no incoming edges is a minimal set of model fragments that can be removed from M. If all the model fragments in C are removable from M, then it is easy to see that $M \setminus C$ is a type 2 simplification of M. Furthermore, one can verify that any type 2 simplification of M must have this form. Finally, it is easy to check that this algorithm runs in polynomial time. These comments make the following lemma immediate:

Lemma 13 *The function find-type-2-simplifications computes the type 2 simplifications of a coherent model in polynomial time.*

The following theorem is an immediate consequence of the above lemmas, and the earlier theorems.

Theorem 6 *Let $\mathcal{I}$ be an instance of the* MINIMAL CAUSAL MODEL *problem such that all the approximation are causal approximations, and the contradictory relation partitions the set $\mathcal{M}$ of model fragments into the set $\mathcal{A}$ of assumption classes. Let each assumption class have a single most accurate model fragment, and let the most accurate model of $\mathcal{I}$ be coherent. Let all the constraints in $\mathcal{C}$ be of the form shown in Equation 20, and let $\mathcal{C}$ contain all the ownership constraints of $\mathcal{I}$. Then a minimal causal model of $\mathcal{I}$ can be found in polynomial time.*

Proof: Immediate from Theorems 3 and 5, and Lemmas 9, 11, and 13 □

5.6 Discussion

The techniques developed in this section have been incorporated into an implemented automated modeling system. This system has been tested on a variety of electromechanical devices drawn from [3; 26; 38], using a library of about 150 types of model fragments, including descriptions of electricity, magnetism, heat, and the kinematics and dynamics of one-dimensional motion. All approximations in this library are, of course, causal approximations. The devices range in complexity from 10 to 54 components, with each device having between 10^{12} and 10^{72} coherent models. In all cases, the system found a minimal causal model in 0.5 to 8 minutes on an Explorer II.

A detailed description of this system and its empirical evaluation are beyond the scope of this paper. The interested reader is referred to [28; 29]. However, two points are worth noting. First, while no ownership constraints were used, there was no loss in solution quality, i.e., minimal causal models were correctly constructed. We conjecture that the reason for this is that most equations describing the physical world do seem to have a natural causal orientation (in which case there are no ownership constraints), and the few situations that do allow multiple causal orientations do not lead to pathological situations. Second, the restricted expressive power of the constraints in $\mathcal{C}$ did not prove to be a limitation. This is because our focus on the task of generating parsimonious causal explanations has made the expected behavior a central criterion for defining model adequacy, thereby decreasing the importance of $\mathcal{C}$ in defining model adequacy.

6 Differential equations

In this section we generalize the results of the previous section to include models involving differential equations. Recall that the treatment in Section 3 focused only on functional dependencies between parameters, and excluded the integration relation between a parameter and its derivative. We represent this integration relation with the *int* equation: $int(\boldsymbol{p}_1, p_2)$ says that p_2 is the derivative of p_1. Note that $int(\boldsymbol{p}_1, p_2)$ can be causally oriented in only one way, to causally determine p_1 by integrating the value of p_2 over time. Given a set E of equations the ***integration completion*** of E makes explicit all such integration links among the parameters of E:

Definition 21 (Integration completion) *Let E be a set of equations. The integration completion of E, denoted $ic(E)$, is defined as follows:*

$$ic(E) = E \cup \{int(\boldsymbol{q}, dq/dt) : dq/dt \in P(E)\}$$

i.e., whenever $P(E)$ contains a derivative, the integration completion of E contains an *int* equation expressing the integration relation. Note that if E contains no differential equations, then $E = ic(E)$.

To include the causal dependency of a parameter on its derivative, we modify Definition 2 to use $ic(E)$ rather than E, i.e., the causal ordering is the transitive closure of an onto causal mapping defined on the integration completion of E. Similarly, in defining complete and overconstrained sets of equations, we modify Definition 3 to use $ic(E)$ instead of E. These modifications are straightforward, and can be found in [28]. We now show how the results of the previous section can also be generalized.

6.1 Canonical form

For many purposes, e.g., numerical integration [8] and causal ordering as defined in [21], sets of differential equations are required to be in *canonical form.* A set of first-order differential equations is in canonical form if each derivative occurs in exactly one equation. For our purposes, we weaken this condition slightly. We shall say that a set of first-order differential equations is in canonical form if each derivative can be causally determined by exactly one equation. Hence, we allow a derivative to occur in more than one equation, though exactly one equation can causally determine it. We enforce this by assuming that the set $\mathcal{M}$ of model fragments is in *canonical form*:

Definition 22 (Canonical form) *A set of model fragments is said to be in canonical form if and only if the following conditions are satisfied:*

1. *All derivatives are local parameters; and*
2. *If derivative dp/dt is local to model fragment m, then dp/dt can be causally determined by exactly one equation in m.*

Condition 1 ensures that derivatives can be determined by the equations of model fragments in exactly one assumption class, while condition 2 ensures that exactly one equation in each such model fragment can determine it. Hence, the above restrictions ensure that the equations of all device models are in canonical form.

A consequence of the above restriction is as follows. Let dp/dt be a derivative that is local to model fragment m, and let e be the equation of m that can causally determine dp/dt. The integration completion of any set of equations that includes e will introduce the equation $int(\mathbf{p}, dp/dt)$. Since dp/dt is local to m, this is *exactly equivalent* to augmenting the equations of m with the equation $int(\mathbf{p}, dp/dt)$. Using this viewpoint, it is easy to verify that if m_2 is a causal approximation of m_1 (without the augmentation), and the same set of derivatives are local to m_1 and m_2, then m_2 remains a causal approximation of m_1 even after the augmentation. This means that, as long as the set of derivatives does not change, all the results of the previous section remain true.

However, it may not always be the case that the same set of derivatives are local to m_1 and m_2. We now discuss this important case.

6.2 Approximating differential equations

Let m_1 and m_2 be model fragments such that m_2 is an approximation of m_1, and let dp/dt be a derivative that is local to m_1, but not local to m_2. Intuitively, m_1

describes a phenomenon using a dynamic model, i.e., a model involving differential equations, while m_2 approximates this description by describing the phenomena using a static, or equilibrium, model. We will consider two types of approximations: called *exogenizing* and *equilibrating* [21]. Exogenizing involves making the assumption that the dynamic behavior of p is slow compared to the time-scale of interest, so that p can be assumed to be constant. Equilibration involves making the assumption that the dynamic behavior of p is much faster than the time-scale of interest, so that p always appears to be in equilibrium ($dp/dt = 0$).

The effect that exogenizing and equilibrating have on a differential equation can be illustrated with the following example. Consider the equation describing the dynamic behavior of an object's temperature:

$$\frac{d\boldsymbol{T}}{d\boldsymbol{t}} = C\boldsymbol{F}$$

where T is the object's temperature, C is its heat capacity, and F is the net heat flow into the object. Exogenizing this equation results in:

$$\mathit{exogenous}(\boldsymbol{T})$$

which states that, at the time-scale of interest, there is no change in the object's temperature. Equilibrating that equation results in:

$$\boldsymbol{F} = 0$$

which states that, at the time-scale of interest, the object's temperature appears to remain in equilibrium with its environment by ensuring that there is no net heat flow into the object. More generally, we have the following definitions of exogenizing and equilibrating:

Definition 23 (Exogenizing and equilibrating) *Let e be a differential equation that can causally determine the derivative dp/dt, i.e., $dp/dt \in P_c(e)$.*

- *Exogenizing e involves replacing it with the equation $exogenous(\boldsymbol{p})$.*
- *Equilibrating e involves replacing it with an equation e' such that (a) $dp/dt \notin P(e')$; (b) $P(e') \subseteq P(e)$; and (c) $P_c(e') \subseteq P_c(e)$.*

Note that, in both exogenizing and equilibrating, the resulting equation does not contain dp/dt.

We will now briefly discuss how the use of differential equations affects the results of the previous section. A careful analysis of those results reveals that the only two results that assumed that models did not contain differential equations were Theorem 4 and Lemma 10. We now show that the former continues to hold, while the latter holds under special conditions.

6.3 Differential equations and the upward failure property

When using differential equations, the proof of Theorem 4 does not apply directly, because of the presence of the additional *int* equations.[12] However, the upward failure property, as embodied in Theorem 4, does continue to hold.

Theorem 7 *Theorem 4 continues to hold even when models include differential equations that can be equilibrated and exogenized.*

Proof: The proof of this theorem is somewhat involved, and we refer the interested reader to [28] for the details. Here we only provide a brief outline of the proof. Let M_1 and M_2 be coherent models such that $M_2 \leq M_1$. Given any onto causal mapping $F_2 : ic(E(M_2)) \rightarrow P(M_2)$, we construct an onto causal mapping $F_1 : ic(E(M_1)) \rightarrow P(M_1)$, such that $C_{F_2} \subseteq tc(C_{F_1})$. Recall that any onto causal mapping can be viewed as a maximum matching in a bipartite graph (see the proof of Lemma 2). Maximum matchings are constructed by first finding an initial, partial matching, and then augmenting this matching using augmenting paths (see [12]). The maximum matching corresponding to F_1 is constructed by using the maximum matching corresponding to F_2 as the initial, partial matching. The crux of the proof lies in showing that each augmentation of the partial matching results in a new matching that, when interpreted as a causal mapping, entails a superset of causal dependencies. A simple induction then shows that the onto causal mapping F_1, corresponding to the resulting maximum matching is such that $C_{F_2} \subseteq tc(C_{F_1})$. □

6.4 Differential equations and immediate simplifications

When models contain differential equations that can be equilibrated, one can show that a coherent model can have an exponential number of immediate simplifications [28]. To address this problem, we impose a restriction on the types of differential equations that can be equilibrated. In particular, we will require that the only equations that can be equilibrated are *self regulating* equations [21].[13] Self regulating equations are differential equations which can causally determine only the derivative, dp/dt, and the parameter, p. Hence, if a self regulating equation is equilibrated, the only parameter that the resulting equation can causally determine is p.

Definition 24 (Self regulating equation) *An equation e is said to be self regulating with respect to the parameter p if and only if $P_c(e) = \{p, dp/dt\}$.*

For example, consider the following self regulating equation, describing the velocity, v, of a falling raindrop with mass m and a coefficient of drag k (g is the acceleration due to gravity) [18]:

[12] If differential equations are not approximated by equilibration, the proof does translate directly (see [28]).

[13] In [28] we identify a slightly more general restriction.

$$m\boldsymbol{dv}/\boldsymbol{dt} = mg - k\boldsymbol{v}$$

The terminal velocity of this raindrop is obtained by equilibrating this equation, resulting in the following:

$$k\boldsymbol{v} = mg$$

Notice that the above equation can only causally determine v. This property is exploited in the following proof of the updated version of Lemma 10.

Lemma 14 *Lemma 10 continues to hold if the only differential equations that can be equilibrated are self regulating equations.*

Proof: The proof is completely analogous to the proof of Lemma 10. The only differences are as follows. First, when constructing $\underline{F}$ from F_2, instead of using Theorem 5, we use an updated version of this theorem, which uses the construction outlined in Theorem 7. Second, F' is defined on the equations in m' using F_2 on the equations in m_2 in the same way, except on differential equations in m' that are approximated in m_2. Let $e' \in m'$ be a differential equation that determines derivative dp/dt, and let $e_2 \in m_2$ be the corresponding approximated equation. Since the model fragments are in canonical form, e' and $int(\boldsymbol{p}, dp/dt)$ must be causally oriented to determine dp/dt and p, respectively. This is well defined extension of the causal mapping F_2 because (a) dp/dt is local to m', and does not occur in m_2; and (b) e_2 must causally determine p. The latter fact follows because if e_2 is an exogenized version of e', then e_2 is $exogenous(\boldsymbol{p})$, and hence e_2 must determine p. On the other hand, if e_2 is an equilibrated version of e', then because e' is a self regulating equation, the only parameter that e_2 can determine is p. In essence, e_2 and e', in conjunction with $int(\boldsymbol{p}, dp/dt)$, behave like corresponding equations. □

The following theorem can be derived from the above two results in much the same way that Theorem 6 was derived:

Theorem 8 *Theorem 6 continues to hold if only self regulating differential equations can be equilibrated.*

7 Related work

One of the original inspirations for the work described here was Davis's work on model-based diagnosis [9]. In that work, Davis presents a diagnostic method based on tracing paths of causal interactions. He argues that the power of the approach stems not from the specific diagnostic method, but from the model which specifies the allowed paths of causal interaction. He shows that efficient diagnosis, while retaining completeness, can be obtained by initially considering models with only a few paths of interactions, and adding in additional paths when the model fails to account for the symptoms.

While we have not focussed on the task of diagnosis, one can see that our simplicity ordering on models lends itself to the above diagnosis technique: diagnosis starts with the minimal causal model, with successively more complex models being used if a model is unable to account for the symptoms. The restrictions in Sections 5 and 6 ensure that using more complex models will add new paths of causal interaction.

The compact representation of the space of device models as a set of model fragments originated in the work on *compositional modeling* [13; 14]. In this work, each model fragment is conditioned on a set of *modeling assumptions*, with mutually contradictory assumptions being organized into assumption classes. A set of constraints govern the use of these assumptions, and a user *query* focuses model selection. An adequate device model is a simplest model that contains all the terms mentioned in the query, and uses only model fragments that are entailed by a set of mutually consistent assumptions satisfying all the constraints. An adequate model is constructed using a variant of constraint satisfaction called *dynamic constraint satisfaction* [27].

The primary difference between their work and ours is in the definition of model adequacy: they have no counterpart of the expected behavior. Our focus on the task of causal explanation has allowed us to use the expected behavior as a central constraint on model adequacy, thereby decreasing the importance of the coherence constraints. This task focus has allowed us to develop a polynomial time algorithm for finding adequate models. Note that the decrease in the importance of coherence constraints means that the restriction on their expressive power (Equation 20) is less serious. On the other hand, in compositional modeling, the constraints on the use of assumptions play a central role in defining model adequacy, and any task focus has to be embedded in these constraints. Embedding such a task focus is, in general, not easy. For example, it is not clear how the expected behavior of a device can be expressed as a set of declarative constraints. Furthermore, any restriction on the expressive power of the constraints would be a serious limitation. Hence, their model selection algorithm is based on dynamic constraint satisfaction, which can, in the worst case, take exponential time.

The definition of model adequacy used in this paper does not explicitly include model accuracy. The work on *graphs of models* [1] discusses a technique for selecting models of acceptable accuracy. A graph of models is a graph in which the nodes are models and the edges are assumptions that have to be changed in moving from one model to another. A model in this graph has acceptable accuracy if its predictions are free of conflicts, which are detected by validating the model's predictions either against empirical data or against consistency constraints. When a conflict is detected, a set of domain-dependent *parameter change* rules help to select a more accurate model, and the above process is repeated. Analysis begins with the simplest model in the graph of models, and terminates when an accurate enough model has been found. Weld extends this work by introducing an interesting class of approximations called *fitting approximations* [42]. Informally, a model M_2 is a fitting approximation of a model M_1 if M_1 contains an exogenous parameter, called a fitting parameter, such that the predictions using M_1 approach the predictions using M_2, as the fitting parameter approaches a limit. Weld shows that when all approximations are fitting approximations, the domain-dependent parameter change rules can be replaced by an efficient domain-independent method for improving model accuracy.

Fitting approximations and causal approximations are fundamentally incomparable because the former talks about behavior differences, while the latter talks about causal dependencies. However, in practice, it appears that fitting approximations are also causal approximations. For example, all the fitting approximations given in [43] are also causal approximations. This means that our model selection method can be combined with his techniques for reasoning about model accuracy, e.g., by using our techniques for selecting the initial model in the graph of models, and using his technique to navigate to models of acceptable accuracy.

In [49], Williams introduces the notion of a *critical abstraction*, which is a parsimonious description of a device relative to a set of questions. Given a device model, he constructs a critical abstraction in three steps: (a) eliminating superfluous interactions; (b) aggregating interactions that are local to a single mechanism using symbolic algebra; and (c) further abstracting the aggregated interactions.

His motivations for creating critical abstractions are very similar to our motivations for finding minimal causal models—we are both striving to find parsimonious descriptions of how a device works. Furthermore, his abstraction process is similar to our model simplification procedure. In fact, the first step of his abstraction process, which eliminates superfluous interactions, is similar to our type 2 simplifications. The primary difference between our approaches is one of emphasis: we have focussed on the problem of selecting approximations from a prespecified space of possible approximations, while he has focussed on finding techniques for automatically abstracting a base model.

8 Conclusions

Constructing adequate problem representations involves the identification of abstractions and approximations that are particularly suited for the problem solving task. In this paper we presented a formalization of the problem of automatically selecting adequate models for physical systems. We formulated this problem as a search problem, requiring answers to the following three questions:

- What is a model, and what is the space of possible models?
- What is an adequate model?
- How do we search the space of possible models for adequate models?

We defined a model as a set of model fragments, where a model fragment is a set of independent algebraic, qualitative, and/or differential equations that partially describes some physical phenomena. The space of possible models was defined implicitly by a library of model fragments: different subsets of model fragments in this library correspond to different models. We gave a precise definition of model adequacy, which was tuned to the task of generating parsimonious causal explanations. An adequate model was defined as a consistent and complete model that could explain the phenomenon of interest. In addition, an adequate model was required to satisfy any applicable domain-dependent constraints. Finally, an adequate model was required to be be as simple as possible, with model simplicity being based on the intuition that modeling fewer phenomena more approximately leads to simpler models.

We then showed that the problem of finding an adequate model is, in general, intractable (NP-hard). In doing this, we identified three different sources of intractability: (a) deciding *what* phenomena to model; (b) deciding *how* to model selected phenomena; and (c) having to satisfy all the domain-dependent constraints.

The intractability of the above problem means that, in general, we can't do much better than search a significant fraction of the whole space of possible models. Unfortunately, even for simple devices, the space of possible models is extremely large, making any sort of brute force search completely impractical. To make model selection practical, we introduced a new class of approximations called causal approximations. Causal approximations form the cornerstone of an important monotonicity property: as models become simpler, the causal relations entailed by the model decrease monotonically. This property allows us to develop an efficient algorithm for finding adequate adequate models. Causal approximations are particularly useful because most commonly used approximations are causal approximations. The results of this paper have been incorporated into an implemented model selection program described in [28; 29].

The work described here can be extended in a number of ways. The most natural extension is to develop more expressive languages for representing the expected behavior. While developing more expressive languages is in itself not difficult, the real challenge is to develop more expressive *tractable* languages. This is important because a central goal of selecting adequate models is to aid effective problem solving. This goal is compromised if the model selection method resulting from using an expressive language for the expected behavior is itself intractable. Hence, an important direction of future research is the development of more expressive languages for expressing the expected behavior that still allow efficient model selection algorithms.

Another natural direction for future work is to develop efficient model selection techniques for other tasks. A particularly promising task appears to be diagnosis, where there is an emerging understanding of what it means for a model to be adequate for diagnosis [9; 20]. Furthermore, as the discussion in Section 7 suggests, we believe that the techniques developed in this paper will prove valuable in developing methods for selecting adequate models for diagnosis.

Acknowledgements

I would like to thank Brian Falkenhainer, Richard Fikes, Andy Golding, Tom Gruber, Yumi Iwasaki, Leo Joskowicz, Rich Keller, Alon Levy, Steve Minton, John Mohammed, and Judea Pearl for helpful discussions on the contents of this paper, and for useful comments on earlier drafts that improved the presentation. Thanks to Jeff Rickel for pointing out an error in an earlier proof of Lemma 8. This work was done at Stanford University, and was supported by an IBM Graduate Technical Fellowship. Additional support for this research was provided by the Defense Advanced Research Projects Agency under NASA Grant NAG 2-581 (under ARPA order number 6822), by NASA under NASA Grant NCC 2-537, and by IBM under agreement number 14780042.

A Examples of causal approximations

In this appendix we present a list of commonly used approximations that can be expressed as causal approximations. Most of these approximations have been borrowed from the fitting approximations listed in [43], though most of the actual equations have been adapted from [18].

Each of the items in this list correspond to a single assumption class. The equations of the different model fragments are presented in a tabular form. Model fragments lower in the table are approximations of model fragments higher in the table, while model fragments at the same level are not approximations of each other.

1. **Translational inertia**

   Newton's second law of motion predicts that the acceleration, a, of a body of mass, m, is proportional to the net force, F, acting on the body. It is common to approximate this law by assuming that the mass, and hence the net force, is zero.

Newton's second law	$\boldsymbol{F} = m\boldsymbol{a}$
No translational inertia	$\boldsymbol{F} = 0$

2. **Rotational inertia**

   This is similar to translational inertia. α is the angular acceleration, I is the moment of inertia, and τ is the net torque.

Newton's second law	$\boldsymbol{\tau} = I\boldsymbol{\alpha}$
No rotational inertia	$\boldsymbol{\tau} = 0$

3. **Relativistic mass**

Einstein's special theory of relativity predicts that the mass, m, of an object increases as its velocity, v, increases. The mass at zero velocity is called the rest mass, m_0. However, this effect is noticeable only at velocities approaching the speed of light, c. At more ordinary velocities, it is common to assume that the mass is constant.

Special theory of relativity	$\boldsymbol{m} = \dfrac{m_0}{\sqrt{1-(v/c)^2}}$
Non-relativistic mass	$exogenous(\boldsymbol{m})$

4. **Relativistic motion**

Let S and S' be observers such that S' is moving at velocity v with respect to S. Let S and S' observe the same event. Let S record the time and position of the event as t and x, and let S' record the time and position of the event as t' and x'. The relationship between x, x', t, and t' is given by the Lorentz transformation. However, at velocities much smaller than the speed of light, c, it is common to use the simpler Galilean transformations.

Lorentz transformation	$\boldsymbol{x}' = \dfrac{\boldsymbol{x} - v\boldsymbol{t}}{\sqrt{1-(v/c)^2}}$ $\boldsymbol{t}' = \dfrac{\boldsymbol{t} - (v/c^2)\boldsymbol{x}}{\sqrt{1-(v/c)^2}}$
Galilean transformation	$\boldsymbol{x}' = \boldsymbol{x} - v\boldsymbol{t}$ $\boldsymbol{t}' = \boldsymbol{t}$

5. **Deformable bodies**

When elastic bodies are acted upon by a force, F, they deform by an amount, x. The deformation is proportional to the force (k is the constant of proportionality), and the relationship between the two is given by Hooke's law. However, it is common to assume that bodies are rigid, so that there is no deformation caused by an applied force.

Hooke's law	$\boldsymbol{F} = -k\boldsymbol{x}$
Rigid bodies	$\boldsymbol{x} = 0$

6. **Friction**

When two bodies move against each other a frictional force, f, impedes the motion. The frictional force is proportional to the force, N, acting normal to the direction of motion, and the constant of proportionality is called the coefficient of friction, μ. However, when motion involves smooth surface, it is common to disregard the frictional force.

Motion with friction	$\boldsymbol{f} = \mu N$
Frictionless motion	$\boldsymbol{f} = 0$

7. **Gravitational fields**

   Newton's law of gravitation predicts that the acceleration due to gravity, g, at a distance r from an object of mass M is proportional to the mass and is inversely proportional to the square of the distance (the constant of proportionality is the Gravitational constant, G). When the variation in r is small compared the magnitude of r, it is common to assume that the acceleration due to gravity is essentially constant. This can be further approximated, when r becomes sufficiently large, by assuming that the acceleration due to gravity is essentially zero.

Newton's law of gravity	$\boldsymbol{g} = GM/r^2$
Constant gravity	$exogenous(\boldsymbol{g})$
Zero gravity	$\boldsymbol{g} = 0$

8. **Collisions**

   Collisions between objects are typically inelastic. If an object approaches a stationary wall at velocity v_i, then the velocity after the collision v_f is attenuated by the coefficient of restitution, α. This is often approximated by assuming that the collision is elastic, so that the initial and final velocities are equal in magnitude.

Inelastic collision	$\boldsymbol{v}_f = -\alpha \boldsymbol{v}_i$
Elastic collision	$\boldsymbol{v}_f = -\boldsymbol{v}_i$

9. **Gas laws**

   The ideal gas law provides a relationship between the pressure, P, the volume, V, and the temperature, T, of a mole of gas. A more accurate gas law is the Van der Waals equation of state, that accounts for the non-zero size of gas molecules, and that gas molecules repel each other at short distances. In these equations, R is the universal gas constant, and a and b are experimental constants.

Van der Waals gas	$(\boldsymbol{P} + \frac{a}{\boldsymbol{V}^2})(\boldsymbol{V} - b) = \boldsymbol{RT}$
Ideal gas law	$\boldsymbol{PV} = \boldsymbol{RT}$

10. **Thermal conduction**

    The rate of heat flow, f, across a thermal conductor is proportional to the difference in temperature at the two ends of the conductor (T_1 and T_2 are the two temperatures). The constant of proportionality is the thermal conductance, γ. There are two different ways of approximating this model. First, we can assume that the conductor is an ideal thermal insulator, so that there is no heat flow. Second, we can assume that the conductor is an ideal thermal conductor, so that there is never a difference between the two temperatures.

Thermal conduction	$\boldsymbol{f} = \gamma(\boldsymbol{T}_2 - \boldsymbol{T}_1)$		
Ideal thermal insulator	$\boldsymbol{f} = 0$	Ideal thermal conductor	$\boldsymbol{T}_1 = \boldsymbol{T}_2$

11. **Thermal conductance**

The thermal conductance, γ, of a thermal conductor is dependent on the length, l, the cross-sectional area, A, and the thermal conductivity, k, of the conductor. When the dependence of γ on these factors is unnecessary, one can merely assume that it is constant.

Dependent thermal conductance	$\gamma = kA/l$
Constant thermal conductance	*exogenous*(γ)

12. **Electrical conduction**

The current flow, i, across an electrical conductor is proportional to the voltage drop, V, across the conductor. The constant of proportionality is the resistance, R, and the relationship is Ohm's law. There are two different ways of approximating this model. First, we can assume that the conductor is an ideal electrical insulator, so that there is no current flow. Second, we can assume that the conductor is an ideal electrical conductor, so that the voltage drop is always zero.

Ohm's law	$\boldsymbol{V} = \boldsymbol{i}\boldsymbol{R}$		
Ideal electrical insulator	$\boldsymbol{i} = 0$	Ideal electrical conductor	$\boldsymbol{V} = 0$

13. **Electrical resistance**

The electrical resistance, R, of an electrical conductor is dependent on the length, l, the cross-sectional area, A, and the resistivity, ρ, of the conductor. When the dependence of R on these factors is unnecessary, one can merely assume that it is constant.

Dependent resistance	$\boldsymbol{R} = \rho l/A$
Constant resistance	*exogenous*$(\boldsymbol{R})$

14. **Resistivity**

The resistivity, ρ, of an electrical conductor is a function of the temperature, T, of the conductor. ρ_0 is the resistivity at temperature T_0, and α is the coefficient of resistivity. However, this dependence is often neglected, and the resistivity is assumed to be constant.

Temperature dependent resistivity	$\rho = \rho_0(1 + \alpha(T - T_0))$
Constant resistivity	$exogenous(\rho)$

15. **Heat engine**

A heat engine can be thought of as a cyclic process that extracts heat from a high temperature source, converts part of this heat into work, and discharges the rest of the heat to a low temperature sink. The efficiency, e, of a heat engine is the fraction of extracted heat that is converted into work. Carnot showed that the efficiency of an ideal heat engine is a function of the source temperature, T_1, and sink temperature, T_2, and that the efficiency of a real heat engine is less than or equal to the ideal efficiency by an efficiency factor, γ.

Real heat engine	$e = \gamma(1 - T_2/T_1)$
Ideal heat engine	$e = (1 - T_2/T_1)$

16. **Laminar flow in horizontal pipes**

The rate, V, of laminar flow of a fluid in a pipe is a proportional to the difference between the pressure at one end of the pipe, P_1, and the pressure at the other end of the pipe, P_2. The pressure drop in the pipe is due to the viscous resistance, R, of the fluid. This model is often approximated to disregard the viscous resistance, so that there is no pressure drop across the pipe.

Viscous flow	$P_1 - P_2 = RV$
Inviscid flow	$P_1 = P_2$

17. **Thermal expansion**

When objects are heated, they expand. The amount of expansion, δ, is a function of the object's temperature, T, and the coefficient of thermal expansion, α. δ is assumed to be zero when the size of the object is l_0 at temperature T_0. This expansion is often quite small, and can be disregarded for many purposes.

Thermal expansion	$\delta = \alpha l_0(T - T_0)$
No thermal expansion	$\delta = 0$

18. **Temperature of an object**

    The rate of change of the temperature, T, of an object is a function of the net heat, F, flowing into the object and the object's heat capacity, C. This equation can be exogenized by assuming that the the temperature is constant, and equilibrated by assuming that the temperature quickly adjusts itself to ensure that the net heat flow is zero. It's not a self regulating equation.

Dynamic thermal model		$\frac{d\boldsymbol{T}}{dt} = C\boldsymbol{F}$	
Constant temperature	$exogenous(\boldsymbol{T})$	Equilibrium temperature	$\boldsymbol{F} = 0$

19. **Falling raindrop**

    The atmospheric drag, mdv/dt, felt by a falling raindrop is proportional to its velocity, v. Eventually, the raindrop reaches its terminal velocity, where dv/dt becomes zero.

Atmospheric drag	$m\frac{d\boldsymbol{v}}{dt} = mg - k\boldsymbol{v}$
Terminal velocity	$mg = k\boldsymbol{v}$

The following are examples of approximations that are not causal approximations.

1. **Viscosity of gases**

   The viscosity, μ, of a gas is a function of its temperature, T, and mass, m, (equivalently, its molecular weight, M). There are at least two models of this dependence. An approximate model assumes that the gas molecules are hard balls of diameter d. A more accurate model models the gas molecule as a force field, and uses the Lennard Jones potential energy function. These models have been taken from [46].

Force field model	$\boldsymbol{\mu} = 2.6693 \times 10^{-6} \frac{\sqrt{MT}}{\sigma^2 \omega_\mu}$
Rigid sphere model	$\boldsymbol{\mu} = \frac{2}{3\pi^{3/2}} \frac{\sqrt{m\kappa T}}{d^2}$

   Note that the force field model does not contain parameters like d that are found in the rigid sphere model. While this approximation does not fit our definition of a causal approximation, one can see that it almost does. In particular, if we are only interested in the dependence of μ on T, then the approximation behaves like a causal approximation.

2. Some approximations are not usually expressed as equations, but are incorporated implicitly into the model. For example, rather than being expressed as an explicit equation, the assumption that a rope is unbreakable is usually incorporated implicitly into the model. This makes it difficult to decide whether or not "unbreakable rope" is a causal approximation.

B Composable operators

Since model fragments are partial descriptions of phenomena, it is useful to allow them to specify partial information about equations. The most commonly used types of partial information, and the ones we consider in this appendix, are the $I+/I-$ operators, first introduced in [15]. These operators allow the specification of partial derivative information. $I+(\boldsymbol{q}_1, q_2)$ states that q_2 is a positive influence on q_1, while $I-(\boldsymbol{q}_1, q_2)$ states that q_2 is a negative influence on q_1. Given a set of positive and negative influences on a parameter, q, a single equation is created by: (a) forming the term resulting from the difference of the sum of the all the positive influences and the sum of all the negative influences; and (b) making the derivative of q be equal to this term. For example, the set $\{I+(\boldsymbol{q}_1, q_2), I-(\boldsymbol{q}_1, q_3), I+(\boldsymbol{q}_1, q_4)\}$ produces the equation (note that only dq_1/dt can be causally determined by this equation):

$$\boldsymbol{dq_1/dt} = q_2 - q_3 + q_4$$

The causal approximation definition can be extended in a straightforward manner to handle $I+/I-$ operators as follows. Let m_2 be a causal approximation of m_1, and let q be a parameter. If m_2 contains influences on q, we will require that the influences on q in m_1 are a superset of the influences on q in m_2. It is then easy to verify that replacing m_2 by m_1 will increase the causal relations monotonically by adding in the new causal relations corresponding to the additional influences. If m_2 does not include influences on q, then we will require that m_1 can contain influences on q only if dq/dt is local to m_1, but not local to m_2. This restriction is completely analogous to Condition 3 in Definition 13, since it ensures that the equation resulting from the newly introduced influences can causally determine dq/dt. To allow influences on a parameter to occur in model fragments of different assumption classes, we need to modify Condition 1 of Definition 22. In particular, derivatives determined by influences will not be required to be local (except when required by the above extension to the causal approximation definition). This modification does not affect the results of Section 6, since the equations resulting from influences cannot be equilibrated or exogenized.

Similar extensions can be made to handle other types of partial information such as the *qualitative proportionalities* ($\alpha_{Q\pm}$) introduced in [15].

References

[1] Sanjaya Addanki, Roberto Cremonini, and J. Scott Penberthy. Graphs of models. *Artificial Intelligence*, 51:145–177, 1991.

[2] Saul Amarel. On representations of problems of reasoning about actions. *Machine Intelligence*, 3:131–171, 1968. Also appears in Webber, Bonnie Lynn and Nilsson, Nils J., editors 1981, *Readings in Artificial Intelligence*. Morgan Kaufmann Publishers. 2–22.

[3] Ivan I. Artobolevsky. *Mechanisms in Modern Engineering Design*, volume 5. Mir Publishers, Moscow, 1980.

[4] Scott W. Bennett. Approximation in mathematical domains. In *Proceedings of the Tenth International Joint Conference on Artificial Intelligence*, pages 239–241, Los Altos, CA, August 1987. International Joint Conferences on Artificial Intelligence, Inc., Morgan Kaufmann Publishers, Inc.

[5] D. Bobrow, editor. *Qualitative Reasoning About Physical Systems.* North-Holland, 1984.

[6] John Seely Brown, R. R. Burton, and Johan de Kleer. Pedagogical, natural language and knowledge engineering techniques in SOPHIE I, II and II. In Derek Sleeman and John Seely Brown, editors, *Intelligent Tutoring Systems*, pages 227–282. Academic Press, New York, 1982.

[7] James Crawford, Adam Farquhar, and Benjamin Kuipers. QPC: A compiler from physical models into qualitative differential equations. In *Proceedings of the Eighth National Conference on Artificial Intelligence*, pages 365–372. American Association for Artificial Intelligence, 1990.

[8] Germund Dahlquist, Åke Björk, and Ned Anderson. *Numerical Methods.* Prentice-Hall, 1974.

[9] Randall Davis. Diagnostic reasoning based on structure and behavior. *Artificial Intelligence*, 24:347–410, 1984.

[10] Johan de Kleer. Multiple representations of knowledge in a mechanics problem-solver. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 299–304. International Joint Conferences on Artificial Intelligence, Inc., 1977.

[11] Johan de Kleer and John Seely Brown. A qualitative physics based on confluences. *Artificial Intelligence*, 24:7–83, 1984.

[12] Shimon Even. *Graph Algorithms.* Computer Science Press, 1979.

[13] Brian Falkenhainer and Kenneth D. Forbus. Setting up large-scale qualitative models. In *Proceedings of the Seventh National Conference on Artificial Intelligence*, pages 301–306. American Association for Artificial Intelligence, 1988.

[14] Brian Falkenhainer and Kenneth D. Forbus. Compositional modeling: Finding the right model for the job. *Artificial Intelligence*, 51:95–143, 1991.

[15] Kenneth D. Forbus. Qualitative process theory. *Artificial Intelligence*, 24:85–168, 1984.

[16] Kenneth D. Forbus. The qualitative process engine. In Daniel S. Weld and Johan de Kleer, editors, *Readings in Qualitative Reasoning about Physical Systems*, pages 220–235. Morgan Kaufmann, 1990.

[17] Kenneth D. Forbus and A. Stevens. Using qualitative simulation to generate explanations. In *Proceedings of the Third Annual Meeting of the Cognitive Science Society*, pages 219–221, 1981.

[18] David Halliday and Robert Resnick. *Physics*. John Wiley and Sons, third edition, 1978.

[19] Walter C. Hamscher. Model-based troubleshooting of digital systems. Technical Report 1074, Artificial Intelligence Laboratory, MIT, Cambridge, MA, 1988.

[20] Walter C. Hamscher. Modeling digital circuits for troubleshooting. *Artificial Intelligence*, 51:223–271, 1991.

[21] Yumi Iwasaki. Causal ordering in a mixed structure. In *Proceedings of the Seventh National Conference on Artificial Intelligence*, pages 313–318. American Association for Artificial Intelligence, August 1988.

[22] Yumi Iwasaki and Chee-Meng Low. Model generation and simulation of device behavior with continuous and discrete changes. Technical Report KSL 91-69, Stanford University, Knowledge Systems Laboratory, 1991.

[23] Yumi Iwasaki and Herbert A. Simon. Causality in device behavior. *Artificial Intelligence*, 29:3–32, 1986.

[24] Benjamin Kuipers. Qualitative simulation. *Artificial Intelligence*, 29:289–338, 1986.

[25] Benjamin Kuipers. Abstraction by time-scale in qualitative simulation. In *Proceedings of the Sixth National Conference on Artificial Intelligence*, pages 621–625. American Association for Artifical Intelligence, AAAI Press, 1987.

[26] David Macaulay. *The Way Things Work*. Houghton Mifflin Company, Boston, 1988.

[27] Sanjay Mittal and Brian Falkenhainer. Dynamic constraint satisfaction. In *Proceedings Eighth National Conference on Artificial Intelligence*, pages 25–32. American Association for Artificial Intelligence, AAAI Press/MIT Press, July 1990.

[28] P. Pandurang Nayak. *Automated Modeling of Physical Systems*. PhD thesis, Stanford University, Department of Computer Science, Stanford, CA, 1992.

[29] P. Pandurang Nayak, Leo Joskowicz, and Sanjaya Addanki. Automated model selection using context-dependent behaviors. In *Proceedings of the Tenth National Conference on Artificial Intelligence*, pages 710–716. American Association for Artificial Intelligence, July 1992.

[30] Ramesh S. Patil, Peter Szolovits, and William B. Schwartz. Causal understanding of patient illness in medical diagnosis. In *Proceedings of the Seventh International Joint Conference on Artificial Intelligence*, pages 893–899. International Joint Conferences on Artificial Intelligence, Inc., 1981.

[31] Olivier Raiman. Order of magnitude reasoning. *Artificial Intelligence*, 51:11–38, 1991.

[32] Earl Sacerdoti. Planning in a hierarchy of abstraction spaces. *Artificial Intelligence*, 5:115–135, 1974.

[33] Elisha Sacks. Piecewise linear reasoning. In *Proceedings of the Sixth National Conference on Artificial Intelligence*, pages 655–659. American Association for Artificial Intelligence, Morgan Kaufmann Publishers, Inc., July 1987.

[34] T. J. Schaefer. The complexity of satisfiability problems. In *Proceedings of the Tenth Annual ACM Symposium on Theory of Computing*, pages 216–226. Association for Computing Machinery, 1978.

[35] Bart Selman and Henry Kautz. Knowledge compilation using horn approximations. In *Proceedings of the Ninth National Conference on Artificial Intelligence*, pages 904–909. American Association for Artificial Intelligence, AAAI Press and The MIT Press, July 1991.

[36] D. Serrano and David C. Gossard. Constraint management in conceptual design. In D. Sriram and R. A. Adey, editors, *Knowledge Based Expert Systems in Engineering: Planning and Design*, pages 211–224. Computational Mechanics Publications, 1987.

[37] Herbert A. Simon. On the definition of the causal relation. *Journal of Philosophy*, 49:517–528, 1952.

[38] C. van Amerongen. *The Way Things Work*. Simon and Schuster, 1967.

[39] J. W. Wallis and E. H. Shortliffe. Explanatory power for medical expert systems: Studies in the representation of causal relationships for clinical consultations. *Methods Inform. Med.*, 21:127–136, 1982.

[40] Sholom M. Weiss, Casimir A. Kulikowski, Saul Amarel, and Aran Safir. A model-based method for computer-aided medical decision-making. *Artificial Intelligence*, 11:145–172, 1978.

[41] Daniel S. Weld. Explaining complex engineered devices. Technical Report TR-5511, BBN, Cambridge, MA, 1983.

[42] Daniel S. Weld. Approximation reformulations. In *Proceedings Eighth National Conference on Artificial Intelligence*, pages 407–412. American Association for Artificial Intelligence, AAAI Press/MIT Press, July 1990.

[43] Daniel S. Weld. Reasoning about model accuracy. *Artificial Intelligence*, 56(2–3):255–300, August 1992.

[44] Daniel S. Weld and Sanjaya Addanki. Query-directed approximation. In Boi Faltings and Peter Struss, editors, *Recent Advances in Qualitative Physics*. MIT Press, Cambridge, MA, 1991.

[45] Daniel S. Weld and Johan de Kleer, editors. *Readings in Qualitative Reasoning About Physical Systems*. Morgan Kaufmann Publishers, Inc., San Mateo, California, 1990.

[46] James R. Welty, Charles E. Wicks, and Robert E. Wilson. *Fundamentals of Momentum, Heat, and Mass Transfer*. John Wiley and Sons, third edition, 1984.

[47] Brian C. Williams. Qualitative analysis of MOS circuits. *Artificial Intelligence*, 24:281–346, 1984.

[48] Brian C. Williams. Interaction-based invention: Designing novel devices from first principles. In *Proceedings Eighth National Conference on Artificial Intelligence*, pages 349–356. American Association for Artificial Intelligence, AAAI Press/MIT Press, July 1990.

[49] Brian C. Williams. Critical abstraction: Generating simplest models for causal explanation. In *Proceedings of the Fifth International Workshop on Qualitative Reasoning about Physical Systems*, May 1991.